Assessing critical thinking in L2: An exploratory study

Sam Reid[1] and Peter Chin[2]
samreid@rikkyo.ac.jp
1. Rikkyo University
2. Waseda University Academic Solutions Corporation
https://doi.org/10.37546/JALTSIG.TEVAL25.1-2

Abstract

Critical thinking (CT) is taking on an increasingly important role in Japanese tertiary education. Teachers tasked with developing CT in a second-language (L2) context may need a way of assessing students' abilities. However, a number of difficulties face L2 students taking a test designed for first-language (L1) speakers. They may be disadvantaged by linguistic and perhaps cultural issues. This study describes an exploratory attempt to make a CT test that can be administered to learners of English and which allows them to display selected elements of CT, specifically analyzing arguments and judging or evaluating. A comparison of L1 and L2 performance in the test showed the results to be comparable. Analysis of two different question topics showed differences in CT skills displayed. Issues with rating accuracy are linked to the format of the test. We argue that this test format is suitable for many students in Japan and elsewhere who have intermediate levels of English.

Keywords: critical thinking, discourse, cognitive load

Critical thinking (CT) may be a feature of tertiary English for Academic Purposes (EAP) courses. Depending on cultural contexts and educational experiences, it may be more or less familiar, and more or less challenging to second language students. It is widely accepted that English as a Foreign Language (EFL) learners require some degree of formal instruction and training in this area (Feng, 2013). This is most especially the case for students going to Anglophone universities. Such institutions often argue that international students do not possess the CT skills necessary for these English-speaking academic cultures (e.g., Fell & Lukianova, 2015; O'Sullivan & Guo, 2011; Shaheen, 2016; Tian & Low, 2011). Furthermore, CT is seen as a requirement to compete in today's global economy (Long, 2004). Despite these academic and financial motivations, there are challenges. In the case of Japan, for example, criticism of students' thinking abilities is a feature of educational discourse (Rear, 2012), and there is growing recognition of "the need for intellectual internationalization and global human resources" (Tsuruta, 2013, p. 147). Considering this, we would agree with Liaw's (2007) view that teachers have a responsibility to help students develop these skills.

CT is a slippery construct. The literature on CT in a first language (L1) differs over issues such as constructs, generalizability, and replication. This uncertainty is heightened when CT is practiced in a second language (L2). Nevertheless, teachers or institutions who make attempts at helping students develop CT need some way of measuring those students' CT ability. This is needed in order to not only assess the current level of their students, but also to track students' progress and measure the effectiveness of CT courses or training within other disciplines that are taught. One option is to use standardized CT tests designed for native English speakers (L1 CT tests). However, as Stroupe (2006) points out, these commercial CT tests may be prohibitively expensive if used on large groups of students. Moreover, as we will argue, using L1 tests for L2 learners may not provide an accurate picture of CT skills, and may be inappropriate for many teaching contexts. An alternative is to convert or translate a test into the students' native language. A drawback to this solution is that it may be expensive and time consuming. Moreover, teachers may also prefer to do a test in English, to serve as more authentic preparation or simulation of study abroad, as well as to simply practice English.

As far as we are aware, no test has been specifically designed to assess CT for L2 students. Therefore, with the aim of ameliorating the disadvantages that L2 students may face, this paper describes an exploratory attempt to make such a test. Our primary concern was whether students taking the test could display equal levels of CT in their L2 as they could in their L1. In addition, we investigated the effect of topic and rating issues. The results indicated that students display similar levels of CT in their L2 and L1. We suggest that the test is flexible and easy to administer, and is particularly suitable for students with intermediate levels of English. We hope it will provide help in the development and guidance of courses meant to foster critical thinking.

Literature review

Although perhaps obvious, it is important to recognize that CT may be more challenging in an L2. A number of studies have looked at the effect of L2 on CT. Davidson and Dunham (1997) used the Ennis-Wier test, in which examinees have to write an evaluation of the arguments in a fictional letter to a newspaper, on tertiary level Japanese students. They found that compared with a control group of 19 students, a treatment group of 17 students who had received instruction in CT did better on the test. Davidson and Dunham's study therefore suggests that it is possible to administer a test designed for L1 to

L2 takers, and that instruction in CT helped them display CT in L2. However, a point to note in this study is that the test conditions did not precisely mirror those for native speakers, as examinees were given twice the standard time to answer the test and were allowed to use dictionaries. In addition, the study does not give any indication of the effect of using CT in L2 rather than L1. Such a comparison is the object of Floyd's (2011) study, in which 55 Chinese students took the Watson-Glaser Critical Thinking Appraisal (Pearson, n.d.). Half of the students took the first half of the test in English and the second half of the test in Chinese, while the other half of the students took the first half of the test in Chinese and the second half in English. Floyd's study used an official licensed translation of the test. The results indicated that displaying CT skills was easier in the L1, a result which was borne out in Floyd's follow-up interviews with participants. There was no time limit and students could use dictionaries, similar to the easing of conditions in the Davidson and Dunham study. A final study addressing this issue is Lun et al. (2010), who were interested in how cultural thinking may affect CT. They administered the Halpern Critical Thinking Assessment using Everyday Situations (Halpern, 2010) and the Watson–Glaser Critical Thinking Appraisal Short Form to students in a New Zealand university. They compared responses from 35 overseas test takers whose L1 was Chinese with those of 24 New Zealand students whose L1 was English and identified as 'New Zealand European'. Results indicated that CT ability was related to general intellectual competence and to English ability, as opposed to cultural thinking styles. In other words, "the difference in critical thinking appears to be more of a linguistic issue rather than a cultural issue" (Lun et al., 2010, p. 613).

There are other studies which use a looser definition of CT, or are not specifically about L2 CT performance, but still shed light on the effect of L2. Luk and Lin (2015) studied a group of Grade 11 students in Hong Kong, and compared what they term 'critical literate talk' in Cantonese and English. Students were tasked with expressing opinions on advertisements, and the definition of CT in this study included generating arguments, evaluating arguments, and making judgments. They found a qualitative difference between the students' ideas expressed in Cantonese and their L2 English, in respect to content and linguistic complexity, and concluded that "The data reveal a wide gap between the students' L1 cognitive maturity and their L2 communicative resources" (2015, p. 70). In terms of written production, Manalo, Watanabe, and Sheppard (2013) investigated university students who wrote about the causes of two disasters, one in their L1 (Japanese) and the other in L2 (English). Their objective was to see if it was harder to be evaluative in Japanese, as Japanese is supposedly less direct in terms of conveying intent or messages. In this study CT was operationalized as students' use of evaluative statements. It is notable again that students were under no time pressure and had received instruction about evaluation. The results showed that students produced more evaluative sentences, evaluative sentences about causes, and evaluative sentences with support when writing in Japanese compared to English. The effect of CT in L2 is shown by "significant correlations between the students' TOEIC scores and their production of evaluative sentences in English (their L2) – but not in Japanese (their L1)" (2013, p. 2971). In other words, their results suggest that although CT is not a linguistic skill, its clear expression requires linguistic ability.

A study by Kaupp et al. (2014) gave three different CT tests to first year students in a Canadian university in an attempt to form a more comprehensive measure of students' CT. Their primary purpose was to assess students' CT development over the course, but as some of the students had English as their L2, they commented on this group when discussing their results. They found that among the three standardized CT tests used – the Cornell Critical Thinking Test: Level Z (Ennis et al., 1985), the International Critical Thinking Essay Test (Paul & Elder, 2010), and the Collegiate Learning Assessment (Council for Aid to Education, n.d.) – only the results from the latter showed significantly lower performance by the English L2 group. Another paper which included a similar analysis of a group of English L2 university students was Facione (1990b), which applied the California Critical Thinking Skills Test (College Level) (Facione, 1990c) to 1,196 students at an American university. Non-native speakers comprised 19% of the sample, and Facione found statistically significant differences between native English speakers and non-native English speakers, who scored lower. His conclusion is unequivocal: "That there is no significant difference from pretest to posttest for non-native English speakers indicates that the CTST instrument is not appropriate for the assessment of college students who are not native English speakers" (1990b, p. 12). The research described so far thus underscores how one must be careful not to mistake a lack of linguistic ability for a lack of CT ability.

Clearly, much of the literature suggests CT is more difficult in L2, so the next issue is why this should be the case. The starting point is the central role of language in CT. According to Moon (2008), although the importance of language differs between CT activities, "it must be seen as extremely important in any critical thinking in the manner that the communication of the thinking is conveyed, distorted, precise or not precise, clear or not clear, subject to manipulation, filled with assumptions, and so on" (p. 73). Similarly, Kobrin et al. (2016) emphasize how language is crucial for both understanding and as a tool for expressing CT. Such views are supported by Takano and Noda (1993), who found that performance in a thinking task declined when a concurrent linguistic task had to be performed in a foreign language. These difficulties are clearly a factor when students take a CT test designed for native speakers. For instance, the Watson-Glaser has been criticized for its unclear instructions and confusing terminology (Possin, 2014). Tellingly, as Kennedy et al. (1991, as cited in Lai, 2011) note, commercially available US CT tests are not designed for students below the fourth-grade level, so tests assume a certain level of linguistic competence that L2 students may not possess. In surveying the constructs which are

assessed by CT tests, Kobrin et al. (2016) note that "Despite differences in the specific knowledge, skills, and abilities measured across critical thinking tests . . ., they all require some verbal ability" (p. 4). This is a particular issue for CT tests which require writing passages. L2 examinees without an advanced level of L2 fluency are bound to make lexical and syntactic errors in their writing and thus may not convey their intended meaning. A rater might disregard such an answer as unacceptable or unclear. Furthermore, L2 examinees may simply decide not to write certain viewpoints because they feel they lack the lexical knowledge to properly explain them in L2. Any attempt to test CT in L2 should take these potential linguistic obstacles into account.

In addition to linguistic knowledge, it has been suggested that differing factual and even cultural knowledge may hamper students. One of the contested points in CT research is about whether CT skills are specific to content areas, or are universal (Moore, 2004). According to Lai (2011), most CT researchers believe background knowledge is important, being "a necessary, though not sufficient, condition for enabling critical thought within a given subject" (p. 42). Norris (1985), for example, argues that successful application of CT requires "among other things, a knowledge of the subject matter, experience in the area in question, and good judgment" (p. 44). To give examples of potential problems with lack of background knowledge, on the California Critical Thinking Skills Test the final four questions refer to a story of a white supremacist and an accompanying scenario in an American school setting, where issues of poverty and race relations arise. In a similar vein, the Ennis-Weir test involves analysis of overnight parking problems. Although this issue is less culturally specific, levels of car ownership and the importance of parking restrictions are not the same in all societies. An important study of whether content familiarity plays a role in critical thinking in relation to L2 writing is Stapleton (2001). In his study of Japanese university students, half the students wrote about rice importation, and half about gun control in the U.S. He found a broader range of arguments and evidence deployed for the familiar topic of rice importation, greater levels of abstraction about the topic, and more references to other viewpoints on the issue. He explains how it is hard to go beyond the literal ideas in the prompt if you do not have background knowledge to tie these ideas to, as wider schema facilitate deeper abstraction about a topic (p. 530). A related study is He and Shi (2012), who tested the effect of topic knowledge on the writing performance of 50 Canadian ESL students with varying degrees of English proficiency. They found that writing performance was better on the general topic of university studies than the specific topic of federal politics. Although this study was not focused on CT, the differences were in "poor idea quality, insufficient idea development, implicit position taking, and weak conclusions" (p. 460), which fall under the scope of CT. Another factor to consider in assessing CT, therefore, is topic choice.

Finally, knowledge may not be a purely factual construct, and may extend to a way, or manner, of thinking. This has important implications for L2 CT, particularly for multiple-choice test formats. Both Ennis (1993, p. 181) and Taube (1995, p. 15) point to how test takers with different assumptions and background beliefs to the test authors' may follow logical lines of reasoning, but will not receive credit for selecting an 'incorrect' answer in tests with a multiple-choice format, where usually no opportunity is given for students to explain the logic behind their selected items. The more distant a student's cultural background, the more likely this becomes. For instance, Fawkes et al. (2005) identified answer choices for the California Critical Thinking Skills Test that potentially have multiple interpretations, thus affecting what can be considered a 'correct' answer. In addition to issues specific to forced choice tests, the vexing topic of 'cultural influence' is relevant to CT more generally. On one side of the debate are Ramanathan and Kaplan (1996), who caution that CT tests examine cultural knowledge that L2 learners may not share, and Atkinson (1997), for whom CT is a social practice better described as cultural thinking. The riposte to these ideas is characterised among others by Davidson (1997), who argues it is more accurate to say that CT is tolerated to different degrees in different spheres of cultures, and Paton (2005), who believes the reasons for student difficulties are lack of practice and topic knowledge rather than thinking styles. With this in mind, raters may need to be aware of such potential differences.

In sum, the difficulties faced by students are neatly summarised by Bali (2015) as "their cultural capital and exposure to critical thinking before college; their exposure to pedagogies that promote critical thinking before college; and their linguistic ability, which impacts their ability to read/write critically" (p. 327). With these things in mind, any attempt to measure CT in a second language has to allow for lower linguistic ability, allow for differences in background assumptions, and allow for differing topic knowledge. This is relevant in the next section, which describes the test format.

Test Format

A number of basic factors were considered important in the context of designing a CT test for an L2 situation. First is the answer requirement. Although a test with a forced choice format can make implementation and rating manageable, as the literature review detailed, this may not be ideal for testing CT. In a synthesis of research on critical thinking, Norris (1985) stated that ideally CT testing requires that takers be productive, not just choose correct options or avoid errors. A second important point is explained by Stroupe (2006), who stressed how assessment of learners' CT in L2 situations must be level appropriate. Lai (2011) advised that "In constructing assessments of critical thinking, educators should use open-ended

tasks, real-world or 'authentic' problem contexts" (p. 42). The final principle is articulated by Facione (1990a), who listed the constructs which should not advantage nor disadvantage students doing a CT test, and among these the important points were reading ability, background knowledge, and culture (p. 32).

A pilot test was carried out in which students were given 20 minutes to write critical responses to the statement "Learning English is necessary for success in today's world". Students had to explain in as much detail as possible why this might not be true. The pilot showed that students required more specific instructions on how to answer, and that students tended to stop writing after about 10 minutes. Based on this, it was decided to give students 10 minutes to write, to provide examples of how to answer, and also to provide two different statements to critically respond to, in order to compare the ease of responding to different topics. Therefore, in the version of the test we trialed, students were given two statements:

- "Learning English is necessary for success in today's world." (henceforth "Learning English")
- "All endangered animals should be saved." (henceforth "Endangered Animals")

These were designed with attention to vocabulary and length, to reduce anything lexically and syntactically challenging to L2 examinees of minimum low-intermediate level. The two topics were chosen based on the authors' attempt to limit possible advantages or disadvantages for students with or without specialist background knowledge. It was presumed that students of diverse cultures would have had sufficient exposure to both issues to be able to articulate views on them. In this way, it was hoped the test was both level appropriate and fair in terms of contextual knowledge. The test was 'productive' in that students had to write as many ideas challenging the statements as possible. Both statements were purposefully vague and easy to challenge, to encourage as wide a variety of responses as possible.

In terms of operationalizing CT for assessment, there is disagreement in the literature as to the scope of what should be tested and the processes that constitute CT. However, as Liaw (2007) argued, there is little essential difference in the various critical thinking definitions. Perhaps the most influential definition of what constitutes CT is Facione's (1990a) consensus statement. This lists the cognitive skills of interpretation, analysis, evaluation, inference, explanation and self-regulation; and the dispositions of critical thinkers, which include open-mindedness regarding divergent world views, flexibility in considering alternatives and opinions, understanding of the opinions of other people, fair-mindedness in appraising reasoning, honesty in facing one's own biases, prejudices, stereotypes, egocentric or sociocentric tendencies, and prudence in suspending, making or altering judgments. This is fairly broad, and for our purposes critical responses were classed as responses which questioned the validity of the original statement, such as counter-arguments, questions about the logic of the statement, or combinations of both. To decide if a response was acceptable, four underlying critical thinking skills were possible:

1. *Seeking clarity* (Is the concept clear? Does any language need to be clarified?)
2. *Challenging the logic* (How true is the statement? Is there any doubt as to its possibility?)
3. *Presenting an alternative viewpoint* (Are there possible negative consequences to consider? Are there more important issues regarding any point in the statement?)
4. *Challenging an assumption* (What are the statement's underlying assumptions? Are these assumptions valid?)

Any response that fulfilled these criteria was classed as acceptable (examples are provided in Appendix A). The more responses the examinees could produce to question the validity of the statements, the better their level of CT. This test therefore focuses on two elements of CT: analyzing arguments, and judging or evaluating. These fit McPeck's (1981, p. 8) definition of CT as "reflective skepticism".

The literature review identified difficulties with L2 and cultural background knowledge as factors which may impact the display of CT. Therefore, when analyzing test performance, we were interested in whether students were disadvantaged by writing in L2. We also wanted to compare the two topics to determine any effect of background knowledge. Finally, as this is an exploratory study, we wanted to see how accurately our operationalization of CT could be judged. With this in mind, the three research questions were as follows:

1. Can participants display evidence of equal CT skills in their L2 as well as they can in their L1?
2. How do the responses produced by the students for each of the two statements compare in terms of number and acceptability?
3. Can the L2 CT test be accurately rated?

Test Implementation

The test was administered at a private Japanese university. The participants comprised 138 students, of whom 102 identified their native language as Japanese, and 36 as Chinese. The students were enrolled in an elective course that focused on critically reading English texts, such as articles and short stories, and discussing their analyses of the readings. The course is recommended for students with a minimum TOEIC 600 or TOEFL iBT 64. However, students were not required to provide proof of TOEIC or TOEFL scores, so could conceivably have been lower, and we did not have measurements of language proficiency. Students took the CT test in the first lesson of the course. All students signed a consent form.

Students had to write as many critical responses to each statement as possible in 10 minutes. Critical responses were single sentences (not paragraphs). For the first statement, students were instructed to write their critical responses in English, and for the second statement in their native language. To illustrate the task requirements and to show students that the L2 linguistic demands on this test are presumably within their range, prior to the start of the test, students were given an example statement ("Dogs make the best pets") and a list of critical responses to the statement (see Appendix B). While writing, students were neither allowed to speak nor use a dictionary.

Two versions of the tests were administered: Practice A and Practice B. In Practice A, students responded to "Learning English" in English, and responded to "Endangered Animals" in their L1. In Practice B students responded to "Endangered Animals" in English and "Learning English" in their L1. Classes were assigned at random to do either Practice A or Practice B. In total, 65 students completed Practice A and 73 completed Practice B.

To ensure accurate rating of the responses written in Japanese and Chinese, we enlisted translations from native speakers. Two raters (the authors of this paper) checked all responses independently and marked each response as acceptable or not. It was agreed in advance that in the case of poor grammar or vocabulary we would give students the benefit of the doubt, following Paul and Elder's (1996, as cited in Stroupe, 2006) suggestion that when assessing CT intellectual standards should be concerned with reasoning over quality of writing. Raters then compared answers. Upon disagreement over acceptability, raters discussed the interpretation of the categories and decided upon a final judgment of acceptable or unacceptable. Examples of acceptable and unacceptable responses can be seen in Appendix A.

Results

The first research question was posed to determine whether participants could display the same level of CT in their first and second languages. Table 1 compares the number of acceptable responses for each statement in participants' L1 and L2. There were more acceptable responses in L1 than in L2 for both statements. The difference was 0.20 more responses in L1 for "Learning English", and 0.31 more responses for "Endangered Animals". The participants in this study included both Japanese and Chinese students, which offered a chance to compare responses by participants' L1. Table 2 breaks down acceptable responses in L1 and L2 by Japanese and Chinese participants. Chinese participants produced a marginally higher mean number of combined L1 and L2 responses for each statement. The difference between the number of L1 and L2 responses was higher among Japanese participants than among Chinese participants. Notably, in the case of Chinese participants responding to "Endangered Animals" there were 0.04 more responses in L2 than L1.

Table 1

Mean acceptable responses for each statement in L1 and L2

Statement	Group	M	SD
Learning English	L1	4.55	2.14
	L2	4.75	2.06
	L1 & L2	4.66	2.09
Endangered Animals	L1	5.97	2.70
	L2	6.26	3.22
	L1 & L2	6.11	2.95

Note: Practice A: n =65; Practice B: n = 73

Table 2

Mean acceptable responses for each statement by participant L1

Participant L1	Statement	Group	M	SD
Japanese	Learning English	L1	4.71	2.15
		L2	4.43	2.15
		L1 & L2	4.6	2.14
	Endangered Animals	L1	6.35	3.45
		L2	5.94	2.64
		L1 & L2	6.10	2.97
Chinese	Learning English	L1	5.00	1.48
		L2	4.76	2.15
		L1 & L2	4.83	1.95
	Endangered Animals	L1	6.12	2.88
		L2	6.18	3.16
		L1 & L2	6.14	2.92

Another possible indication of whether participants could display equal CT in L1 and L2 was the number of responses written in each language that were rated as acceptable. If expressing ideas is more difficult in L2, there may be fewer acceptable responses in L2 compared with L1. Table 3 shows the percentage of responses graded as acceptable, and whether responses were in L1 or L2. For both statements, more responses were graded as acceptable in L1. Taking L1 and L2 together, 14.74% more responses were rated as acceptable for "Endangered Animals" than "Learning English".

Table 3

Percentages of responses graded acceptable for each statement by participant L1

Statement	Group	Acceptable (%)
Learning English	L1	71.99
	L2	64.77
	L1 & L2	68.48
Endangered Animals	L1	85.32
	L2	81.94
	L1 & L2	83.22

Note: Practice A: n =65; Practice B: n = 73

The second research question was posed to compare the number and acceptability of responses to each statement. Table 1, above, shows that the mean responses to "Learning English" was 4.66 and the mean responses to "Endangered Animals" was 6.11, and so 1.45 greater for "Endangered Animals". In order to add detail to this, Table 4 shows the mean number of acceptable responses per participant, as well as the overall number of responses to each statement, and the variety of different ideas given for each statement (in other words, different possible responses). As well as having more overall responses, "Endangered Animals" had a greater variety of responses in terms of content. This shows that there were more ideas produced in response to "Endangered Animals".

Table 4

Number of acceptable responses and variety of acceptable ideas for each statement

	Learning English	Endangered Animals
Overall number of responses	643	843
Variety of ideas	113	135

Figures 1 and 2 are box plots comparing the mean number of responses to each statement broken down into nationality and L1/L2. They show how "Endangered Animals" was easier to respond to compared with "Learning English", as indicated by inter-quartile spreads as well as bottom and top whisker lengths. In Figure 1, among the Japanese L1 participants, boxplot spread, skew, and median value were identical regardless of L1 or L2 use. In Figure 2, there were differences when it came to responding to "Endangered Animals", with a wider range of responses in L2. The spread of responses among the Chinese participants were more compact compared to the Japanese participants.

The third research question concerned whether the test could be accurately rated. There were two stages to the grading process. To begin with, graders separately checked all responses and decided upon acceptability. Second, graders came together to compare results and discuss cases of disagreement over acceptability to decide these cases together. When raters compared together after the first separate check, inter-rater agreement of acceptable responses to "Learning English" was 84.96%, and agreement of acceptable responses to 'Endangered Animals" was higher at 92.59%. For both statements combined it was 89.11%.

Figure 1

Acceptable responses to "Learning English" by L1 and L2

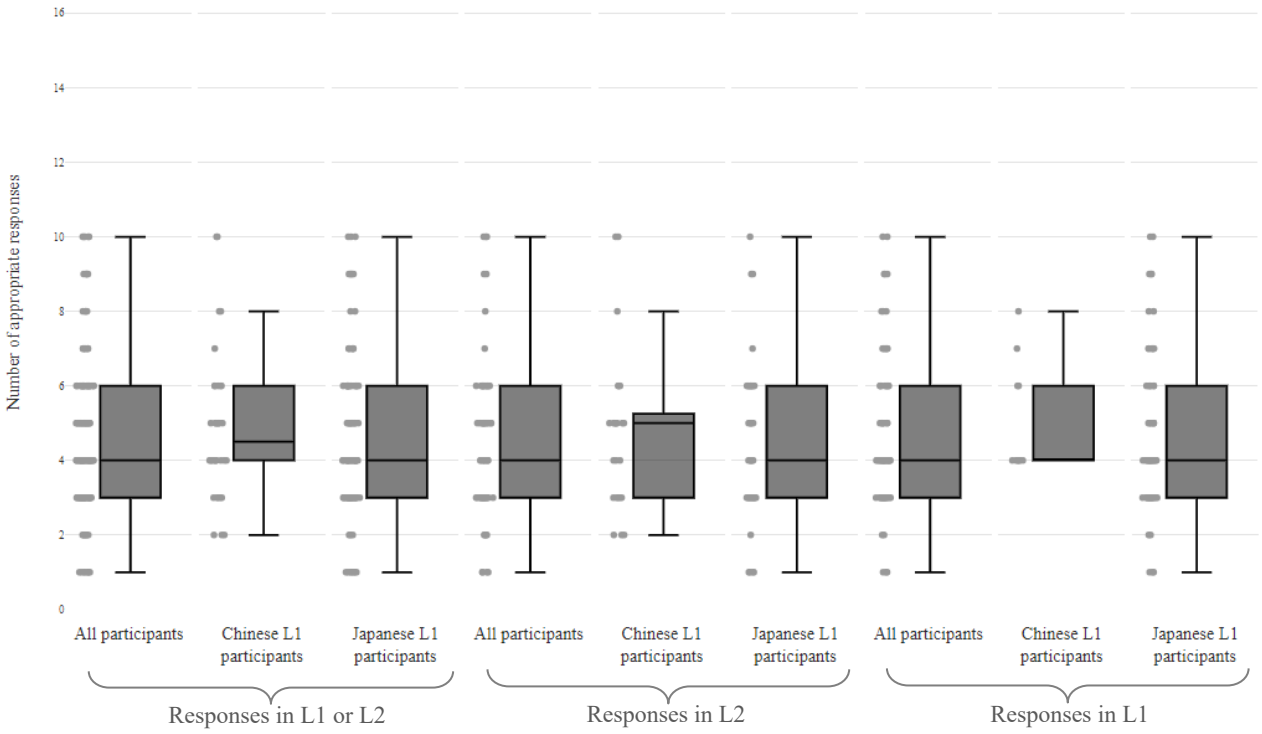
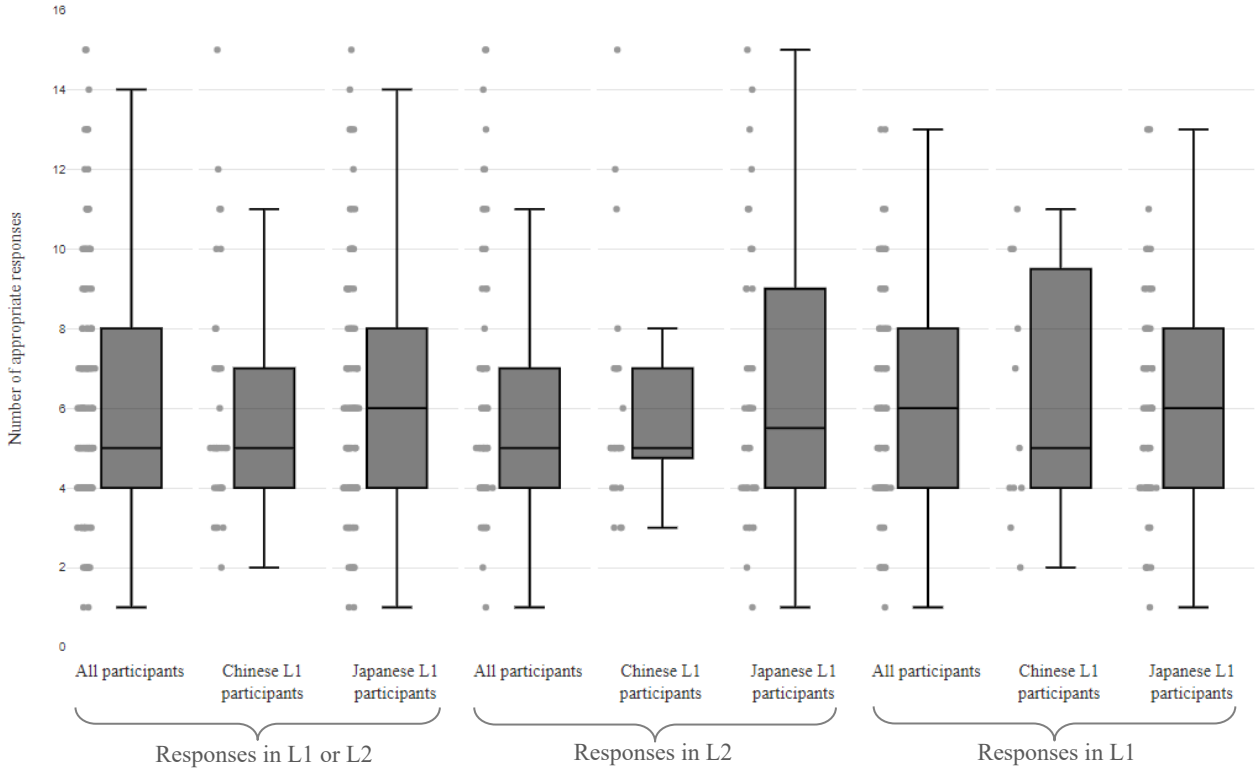


Figure 2

Acceptable responses to "Endangered Animals" by L1 and L2



Discussion

To recap, the research questions concerned the level of CT participants could display in L1 and L2, the comparative difficulty of the two statements, and the accuracy of rating.

Research Question 1: Can participants display evidence of equal CT skills in their L2 as well as they can in their L1?

Participants performed slightly better in their L1 than their L2: The difference in mean acceptable L2 responses were 4.55 to 4.75 for "Learning English", and 5.97 to 6.26 L1 for "Endangered Animals". Therefore, the goal of creating a test which did not disadvantage L2 participants was not completely successful, even though the differences between L1 and L2 responses do not appear extreme. Overall, then, this study supports the idea noted elsewhere that displaying CT skills is more difficult in a L2 due to linguistic difficulties, rather than deficient CT ability. However, the fact that Chinese students responding to "Endangered Animals" produced a slightly higher mean of acceptable L2 responses than in L1 suggests that language may not always be an inhibiting factor. The study showed slightly higher CT levels for the Chinese students. This could be because students who have the aptitude and resources to study abroad (as is the case with Chinese students studying in Japan) possibly have higher academic levels and/or language proficiency.

As well as a higher mean of acceptable responses, the results showed a slightly higher percentage of responses were rated as acceptable in L1 than L2. Interestingly, however, language comprehensibility was not an issue. All but three of the responses written in L2 were comprehensible for the raters. Answers were overwhelmingly deemed unacceptable through weak CT rather than lack of comprehensibility. As there were 992 total responses in English, to have only three responses (0.30%) determined to be incomprehensible due to issues with L2 lends support to the format of the test being appropriate for L2 students at a proficiency level of around TOEIC 600 and over. It is difficult to ascertain the reason for this greater response acceptability in L1. One possibility is that participants felt more confident expressing ideas in L1 and hesitated to express ideas in L2.

A drawback of this test format is the limited operationalization of CT in the test. In making a test that is suitable for students who cannot produce longer written passages certain compromises were necessary, one of which was that the test focused on deconstructing rather than constructing arguments. It does not measure other possible CT aspects, such as the ability to argue for a position, using supporting reasons and examples, judging evidence, or deciding on a course of action, all of which have been included in definitions of CT. Other tests that have been found to disadvantage L2 participants more than this one may measure a more comprehensive operationalization of CT. Another drawback is that L2 ability has been identified as a factor in display of CT in L2, but we did not take L2 ability into account. Regrettably, we did not have a standardized test measure to factor in, such as a TOEIC or TOEFL score. Students on these courses were expected to have an English level of intermediate and above, but we lacked the means to differentiate students based on English levels.

Research Question 2: How do the responses produced by the students for each of the two statements compare in terms of number and acceptability of responses?

The results clearly show that "Endangered Animals" was easier to respond to than "Learning English". Mean responses were higher, the percentage of responses rated as acceptable was higher, and the variety of ideas rated as acceptable was higher. The higher inter-rater agreement about responses to "Endangered Animals" may also suggest that this was easier to respond to. "Learning English" may be a more complex issue in terms of critical responses for a number of reasons. Perhaps students' first-hand experience with this topic made responding critically to long-held assumptions about it more formidable, or contributed to students drawing more on illogical or fallacious ideas than when responding to "Endangered Animals". Regardless of the reason, the present study lends support to the finding that topic affects CT display. The fact that the range of responses differed more for "Endangered Animals" might indicate that the more open to criticism a statement is, the more the use of L1 or L2 can impact performance. If this is true, it would affect the degree to which the test is able to differentiate high from low performers, both regarding CT ability (number of acceptable responses) and strength of L2 ability (in comparison to L1 results).

Further research would help clarify which topics may be most appropriate for general or particular kinds of English L2 students. For instance, other statements could be "Famous people are good role models because they are successful" or "It is important to protect the natural environment for future generations", as it would be assumed most cultures have celebrities and protecting the environment is a universally debated topic. In addition, future research could involve interviews with participants to ask whether one topic was more difficult than the other and why.

Research Question 3: Can the L2 CT test be accurately rated?

With respect to rating, we consider 89.11% initial agreement on how to classify responses an acceptable level, considering the open-ended format of the test, potential variability in interpretations of criteria, and differences in the cross-cultural backgrounds of those involved in the test (both students' and raters'). All differences were resolved through rater discussion, which served to further refine the criteria. Rating difficulties is a complex topic that merits more discussion.

One source of disagreement was about the scope of the categories. For instance, two responses to "Learning English" were *Might produce a world centering only on U.S.* and *Lead to a lack of study of native language*. One rater considered these to be unacceptable because they are not problems with the logic or feasibility of the statement itself, but arguments against "Learning English". However, after discussion, it was decided to include criticisms based on unintended or undesirable consequences. A further example concerns the following responses to "Endangered Animals": *It is decided by God* and *Eating whale is a Japanese tradition.* Disagreement centered on whether to allow for possible adherence to religion or traditions (which are not CT in the sense of analytical logic). For instance, in some cultures, it is believed that nature is controlled by God, and that upholding tradition is regarded as more important than protecting endangered animals. These responses were eventually deemed acceptable in consideration of what might have been positions based on the students' cultural or personal beliefs.

The second type of problem was with the level of detail required for an acceptable response. This area is arguably more complex. To illustrate, the following responses were frequently given for both statements: *It takes too much time* and *It costs too much money.* These answers seem to imply the following: "The time and money which would be spent on these endeavors would be better used for other purposes". Again, the issue comes up of differing background beliefs. We assumed that students felt that this was shared context that did not require further explanation, and decided to accept such responses as acceptable. However, the inferences involved in other responses were less clear. For example, two responses to "Learning English" were *Nationality is more important* and *It reduces your chances to learn other languages.* Similarly, responses to "Endangered Animals" were *Humans are animals too, and it is selfish to regard humans as different from other animals* and *People may use them to make money.* It was decided that such responses could have been acceptable had the students explained their thoughts further. However, at some point the potential reasons were numerous or not immediately obvious, and it was at this point that an answer became unacceptable.

This issue of category boundaries may be resolvable with clearer pre-rating guidelines. For example, a list of standardized acceptable and unacceptable answers given to students before the test would be useful. However, the scope of acceptable answers is a more serious issue. The rating examples highlight an intrinsic difficulty with the format of the test (and perhaps any CT test based on assessing production), which is the subjective nature of the assessment criteria. There is a need for a cut-off point between what is inferable (and thus acceptable), and what requires further explanation (and thus is not acceptable).

Another limitation of this test format is the ambiguity of deciding the cut off point for what constitutes enough support or explanation for an idea. We did not pre-discuss the implementation of categories because we wanted to see what issues would become apparent. The difficulty in this study with identifying CT aligns with what has been noted by others. For instance, in rating L2 essays for evidence of CT skills Stapleton (2001) commented that agreeing on categorization was difficult, and agreeing on what was acceptable or not acceptable was also difficult. He also noted that while there is discussion of the idea of critical thinking itself, there are few criteria or scoring guides. In other words, it is easier to provide an abstract definition of CT than to delineate a cut-off point between a concrete phrasing that is and is not sufficiently critical. On a similar note, Possin (2014) highlights the difficulty in how judgments about the acceptability of conclusions may be different between test takers and test raters because of differing background beliefs. Finally, Norris (1989) stresses that while any answer key represents a test maker's judgment of what is acceptable, the "test maker must take into account … background empirical beliefs and political and religious ideologies that reasonably could be expected to be held by examinees, and assumptions that examinees would likely make" (p. 23). These issues came to the fore when rating, and it appears cultural differences in what constitutes shared knowledge and CT expectations contributed to some statements being deemed unacceptable. The ongoing debate over definitions of CT suggest that it is also an issue in L1 assessment, not just L2 assessment.

Conclusions

With this CT test format, participants made more responses in their L1 than in their L2. Also, the topic that participants responded to was an important factor in how much CT they could display. Finally, issues for grading were not related to understanding student responses in terms of the linguistic content, but about the boundaries of what constitutes CT. Overall, our study supports previous findings concerning the increased difficulty of CT tasks when performed in the L2. However, we do not feel that the small differences observed between the L1 and L2 performance negate the usefulness of the test. The test format is quick to implement and is easily adaptable in terms of the statements to respond to. Such a format may point the way to an appropriate testing instrument for the many instructors in Japan and other countries with students who do not have a suitable level of English for an L1 test, or who have not been trained in formal writing in the L2. We encourage others to modify and refine this format.

Acknowledgements

The authors are very grateful for the detailed feedback provided by the reviewers and editor.

References

Atkinson, D. (1997). A critical approach to critical thinking in TESOL. *TESOL Quarterly, 31*(1), 71-94. https://doi.org/10.2307/3587975

Bali, M. (2015). Critical thinking through a multicultural lens: Cultural challenges of teaching critical thinking. In M. Davies, and R. Barnett (Eds.), *The Palgrave Handbook of Critical Thinking in Higher Education* (pp. 317-334). Macmillan.

Council for Aid to Education. (n.d.). *Collegiate Learning Assessment*. https://cae.org/

Davidson, B. W. (1997). Comments on Dwight Atkinson's "A critical approach to critical thinking in TESOL": A case for critical thinking in the English language classroom. *TESOL Quarterly, 32*(1), 119-123. https://focionline.files.wordpress.com/2014/05/atkinson-comment-1.pdf

Davidson, B. W., & Dunham, R. A. (1997). Assessing EFL student progress in critical thinking with the Ennis-Weir Critical Thinking Essay Test. *JALT Journal, 19*(1), 43-57. http://jalt-publications.org/jj/articles/2704-assessing-efl-student-progress-critical-thinking-ennis-weir-critical-thinking-essay

Ennis, R. H., Millman, J., & Tomko, T. N. (1985). *Cornell Critical Thinking Tests Level X & Level Z: Manual.* Midwest Publications.

Ennis, R. H. (1993). Critical Thinking Assessment. *Theory into Practice, 32*(3), 179-186.

https://doi.org/10.1080/00405849309543594

Facione, P. A. (1990a). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*. https://files.eric.ed.gov/fulltext/ED315423.pdf

Facione, P. A. (1990b). The California Critical Thinking Skills Test-College Level. Technical Report #2. Factors Predictive of CT Skills. California Academic Press.

Facione P. A. (1990c). *California Critical Thinking Skills Test: College Level.* California Academic Press.

Fawkes, D., O'Meara, B., Weber, D., & Flage, D. (2005). Examining the exam: A critical look at The California Critical Thinking Skills Test. *Science & Education, 1*(4), 117-135. https://doi.org/10.1007/s11191-005-6181-4

Fell, E. V., & Lukianova, N. (2015). British Universities: International students' alleged lack of critical thinking. *Procedia – Social and Behavioural Science, 215*(8), 2-8. https://doi.org/10.1016/j.sbspro.2015.11.565

Feng, Z. (2013). Using teacher questions to enhance EFL students' critical thinking ability. *Journal of Curriculum and Teaching, 2*(20), 147-153. https://doi.org/10.5430/jct.v2n2p147

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

Halpern, D. F. (2010). *Halpern Critical Thinking Assessment.* Schuhfried.

He, L., & Shi, L. (2012). Topical knowledge and ESL writing. *Language Testing, 29*(3), 443-464. https://doi.org/10.1177/0265532212436659

Kaupp, J., Frank, B., & Chen, A. (2014). *Evaluating critical thinking and problem solving in large classes: Model eliciting activities for critical thinking development*. Higher Education Quality Council of Ontario. http://www.heqco.ca/SiteCollectionDocuments/Formatted%20Queen%27s_Frank.pdf

Kobrin, J. L., Sato, E., Lai, E., & Weegar, J. (2016, April 9-11). *Examination of the constructs assessed by published tests of critical thinking* [Paper Presentation]. Annual Meeting of the National Council on Measurement in Education, Washington, D.C.

Lai, E. R. (2011). *Critical thinking: A literature review*. Pearson. http://images.pearsonassessments.com/images/tmrs/CriticalThinkingReviewFINAL.pdf

Liaw, M. (2007). Content-based reading and writing for critical thinking skills in an EFL context. *English Teaching & Learning, 31*(2), 45-87. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.580.288&rep=rep1&type=pdf

Long, C. C. (2004). *Teaching Critical Thinking in Asian EFL Contexts: Theoretical issues and Practical Applications*. www.paaljapan.org/resources/proceedings/PAAL8/pdf/pdf022.pdf
http://paaljapan.org/resources/proceedings/PAAL8/pdf/pdf022.pdf

Luk, J., & Lin, A. (2015). Voices without words: Doing critical literate talk in English as a second language. *TESOL Quarterly, 49*(1), 67-91. https://doi.org/10.1002/tesq.161

Lun, V. M., Fischer, R., & Ward, C. (2010). Exploring cultural differences in critical thinking: Is it about my thinking style or the language I speak? *Learning and Individual Differences, 20*, 604–616. https://doi.org/10.1016/j.lindif.2010.07.001

Manalo, E., Watanabe, K., & Sheppard, C. (2013). Do Language Structure or Language Proficiency Affect Critical Evaluation? *Proceedings of the Annual Meeting of the Cognitive Science Society, 35*(35), 1069-7977.

McPeck, J. E. (1981). *Critical thinking and education.* St. Martin's Press.

Moon, J. (2008). *Critical thinking: An exploration of theory and practice.* Routledge.

Moore, T. (2004). The critical thinking debate: How general are general thinking skills? *Higher Education Research & Development, 23*(1), 3-18. https://doi.org/10.1080/0729436032000168469

Norris. S. P. (1985). Synthesis of research on critical thinking. *Educational Leadership*, 40-45. http://www.ascd.org/ASCD/pdf/journals/ed_lead/el_198505_norris.pdf

Norris, S. P. (1989). Can we test validity for critical thinking? *Educational Researcher, 18*(9), 21-26. https://doi.org/10.3102/0013189X018009021

O' Sullivan, M. W., & Guo, L. (2011). Critical thinking and Chinese international students: An East-West dialogue. *Journal of Contemporary Issues in Education, 5*(2), 53-73. http://dx.doi.org/10.20355/C5NK5Z

Paton, M. (2005). Is critical analysis foreign to Chinese students? In E. Manalo. & G. Wong-Toi (Eds.), *Communication skills in university education: The international dimension* (pp. 1–11). Pearson Education.

Paul, R., & Elder, L. (2010). *International Critical Thinking Test.* Foundation for Critical Thinking.

Pearson. (n.d.). *Watson Glaser Critical Thinking Appraisal*. https://www.talentlens.co.uk/product/watson-glaser/

Possin, K. (2014). Critique of the Watson-Glaser Critical Thinking Appraisal Test: The more you know, the lower your score. *Informal Logic, 34*(4), 393-416. https://doi.org/10.22329/il.v34i4.4141

Ramanathan, V., & Kaplan, R. B. (1996). Some problematic "channels" in the teaching of critical thinking in current LI composition textbooks: Implications for L2 student-writers. *Issues in Applied Linguistics, 7*(2), 225-249.

Rear, D. (2012). The dilemma of critical thinking: conformism and non-conformism in Japanese education policy. In T. Isles & P. Matanle (Eds.), *Researching Twenty-First Century Japan: New Perspectives for the Electronic Age* (pp. 119 – 137). Lexington Books.

Shaheen, N. (2016). International students' critical thinking–related problem areas: UK university teachers' perspectives. *Journal of Research in International Education, 15*(1), 18-31. https://doi.org/10.1177/1475240916635895

Stapleton, P. (2001). Assessing critical thinking in the writing of Japanese university students: Insights about assumptions and content familiarity. *Written Communication, 18*(4), 506-548. https://doi.org/10.1177/0741088301018004004

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

Takano, Y., & Noda, A. (1993). A temporary decline of thinking ability during foreign language processing. *Journal of Cross-Cultural Psychology, 24*(4), 445-462. https://doi.org/10.1177/0022022193244005

Taube, K. T. (1995, April 18-22). *Critical thinking ability and disposition as factors of performance on a written critical thinking test*. [Paper presentation] Annual Meeting of the American Educational Research Association, San Francisco, CA.

Tian, J., & Low, G. (2011). Critical thinking and Chinese university students: A review of the evidence. *Language, Culture and Curriculum, 24*(1), 61–76. https://doi.org/10.1080/07908318.2010.546400

Tsuruta, Y. (2013). The knowledge society and the internationalization of Japanese higher education. *Asia Pacific Journal of Education, 33*(2), 140–155. http://dx.doi.org/10.1080/02188791.2013.780674

Appendix A

Sample acceptable critical responses

	Learning English	Endangered Animals
Seeking clarity	・ How much "learning" is enough? ・ "Success" has different meanings to different people.	・ What is the definition of "saved"? ・ Why all?
Challenging the logic	・ Just learning English will not lead to success. ・ If you are already successful in your own career, you don't need to learn English.	・ It's impossible to save and take care of all endangered animals. ・ If there's only one left of an animal, we can't do anything.
Presenting an alternative viewpoint	・ People who can speak English can help those who can't. ・ Getting knowledge is more important.	・ New animals can replace them. ・ To save children in developing countries is more important than saving endangered animals.
Challenging an assumption	・ Lots of people do not speak English but are successful/rich. ・ Not all jobs require English.	・ Not everyone cares about animals. ・ Animals probably aren't happy under our protection.

Appendix B

Instructions given to students

Model presented in both Practice A and B:

S16-

Critical Thinking Example

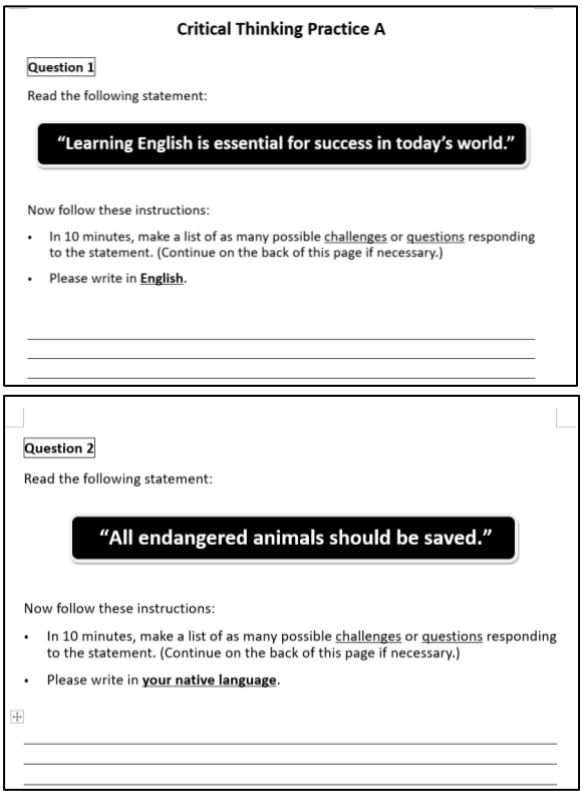
Read the following statement:

"Dogs make the best pets since they are loyal and friendly."

Below is a list of challenges or questions in response to this statement:

- *What do you mean by "loyal?"*
- *Cats can clean themselves, so they don't usually smell. However, you have to give a dog a bath, which is time consuming and can be messy.*
- *You need to walk a dog twice a day. If you're sick or don't like going outside, that can be a problem.*
- *Dogs bark loudly. This will likely disturb your neighbors.*
- *Hairy dogs shed hair. You might develop an allergy from that.*
- *Dogs need a lot of space to run around. If you have a small apartment, your dog may not be happy.*
- *Fish are better than dogs since they require less space.*
- *Some dogs are dangerous and will attack or even kill people.*
- *Could you explain "friendly" more?*
- *Dog food can be expensive. Fish food is much cheaper than dog food.*
- *If the dog is sick, you have to take it to the hospital, so medical expenses for a dog can be expensive.*
- *It'll be more difficult to find an apartment because many apartment owners won't rent to dog owners.*
- *Dogs may not want to be kept as a pet. They might feel lonely when you're not around. Also, most dogs can't pee or take a dump whenever they want because they have to wait for you to take it out for a walk.*
- *Is it really good to keep a pet, like a dog? If you have really strong feelings for your pet, you might suffer mentally when it dies.*
- *Is there any evidence dogs are more loyal than other pets, like a monkey or a cat?*
- *How do we know dogs are really loyal and friendly? Maybe they just want food.*

Practice A instructions:



Critical Thinking Practice A

Question 1

Read the following statement:

"Learning English is essential for success in today's world."

Now follow these instructions:

- In 10 minutes, make a list of as many possible challenges or questions responding to the statement. (Continue on the back of this page if necessary.)
- Please write in **English**.

Question 2

Read the following statement:

"All endangered animals should be saved."

Now follow these instructions:

- In 10 minutes, make a list of as many possible challenges or questions responding to the statement. (Continue on the back of this page if necessary.)
- Please write in **your native language**.

*Statements reversed in Practice B.