RESEARCH Open Access

Assessing critical thinking through L2 argumentative essays: an investigation of relevant and salient criteria from raters' perspectives

Takanori Sato

Correspondence: taka-sato@sophia.ac.jp
Sophia University, Bld#6 510, 7-1
Kioi-cho, Chiyoda-ku, Tokyo
102-8554, Japan

Abstract

Although some second language (L2) pedagogical approaches recognize critical thinking (CT) as an important skill, its assessment is challenging because it is not a well-defined construct with varying definitions. This study aimed to identify the relevant and salient features of argumentative essays that allow for the assessment of L2 students' CT skills. This study implemented a convergent mixed-methods research design, collecting and analyzing both quantitative and qualitative data to collate the results. Five raters assessed 140 causal argumentative essays written by Japanese university students attending Content and Language Integrated Learning courses based on five criteria: Task Achievement, Coherence and Cohesion, Lexical Resource, Grammatical Range and Accuracy, and CT Skills. A standard multiple regression was conducted to examine the relationships among these criteria. Additionally, raters' written verbal protocols were collected to identify the essay features to be considered when assessing students' CT skills. The results indicated that raters' judgments of students' CT were closely linked to Task Achievement. Furthermore, their assessments were affected by the essay's relevancy to the question, content development, logicality, and quality of ideas. This study's findings help to conceptualize CT as a construct and should be incorporated into the assessment criteria of various L2 educational contexts.

Keywords: Critical thinking, Second language writing, Argumentative writing, Assessment criteria, Rating scale development

Introduction

Some second language (L2) pedagogical approaches, including English for academic purposes (EAP) and Content and Language Integrated Learning (CLIL), stress the importance of critical thinking (CT) since the skills are vital in academia and help students engage with world knowledge (de Chazal, 2014; Mehisto & Ting, 2017). Being an integral part of such instructional approaches, the assessment of CT must be conducted to foster decisions on summative and formative purposes in the course. In this context, essay writing assignments are considered as an effective tool for assessing CT skills, as they provide students with time to carefully consider reasons for their assertions and refine their ideas (Nosich, 2022; Wade, 1995).

However, assessing CT through essay writing is challenging because "the term 'critical thinking' is a notoriously fuzzy construct in education" (Yuan & Stapleton, 2020, p. 41) and "critical thinking as a concept is diffuse" (Wilson, 2016, p. 257). While performance assessment requires rating scales to enable assessors to measure students' L2 output (McNamara, 1996), the elusive construct of CT makes it difficult to clearly decide what to assess. Although numerous elements of CT have been explicated for general education (e.g., Paul & Elder, 2014), it has not been operationalized specifically for L2 pedagogical settings, and hence relevant and salient criteria have not been established for assessing L2 students' CT through their essays. In particular, delineating this construct is warranted for argumentative writing, which is an imperative type of writing that L2 students are likely to engage in various academic contexts (Hirvela, 2017).

One effective approach to disentangling such an elusive construct is to investigate people's intuitive judgments of it. In other words, studies on how essay readers evaluate the writers' CT skills can provide empirical data that can help researchers identify relevant and salient features of the construct. Nevertheless, no existing studies have implemented this research approach to delineate CT for L2 writing assessment. Therefore, this study aimed to identify the features of argumentative essays that allow for the assessment of L2 students' CT by investigating how readers rate and judge the writers' CT skills through their argumentative essays. This study's findings contribute to the conceptualization of CT as a construct and the development of rating scales for measuring it in L2 educational contexts.

Literature review

CT theories and argumentative writing

CT is known as a fuzzy and elusive concept because of its various competing definitions and interpretations (Wilson, 2016). Davies and Barnett (2015) indicate how widely CT has been defined by summarizing its concepts in three movements: the critical thinking movement, which focuses on argumentation skills and dispositions; the criticality movement, which addresses ethical actions and morality in society; and the critical pedagogy movement, which aims to overcome the oppression that restricts human freedom. One widely utilized definition for CT is "Critical thinking is a reasonable reflective thinking focused on deciding what to believe and do" (Ennis, 2011, p. 10). This conception outlines 12 dispositions (e.g., trying to be well informed and being open-minded) and 16 abilities (e.g., analyzing arguments and judging the credibility of sources) that describe the characteristics of ideal critical thinkers. While argumentative skills are required to demonstrate reflective thinking, this view focuses on judgment formation and decision-making than the mechanisms of argumentation (Davies & Barnett, 2015). Ennis (2011) claims that CT "should be a very important part of our personal, civic, and vocational lives and should receive attention in our education system" (p. 10).

CT skills in L2 pedagogies are built on Ennis's (2011) conception and focus on promoting argumentation and cognitive thinking skills. Dummett and Hughes (2019) defined CT in the English language teaching context as "a mindset that involves thinking *reflectively* (being curious), *rationally* (thinking analytically), and *reasonably* (coming to sensible conclusions)" [emphasis in original] (p. 4) and illustrated how it is associated with Anderson and Krathwohl's (2001) categories of cognitive process dimensions. Anderson and Krathwohl (2001) specified six cognitive process categories that education should incorporate to help students improve their retention abilities and the transfer of learning. These categories are as follows: to remember (retrieving knowledge from memory), understand (building connections between prior and new knowledge), apply (using the acquired knowledge in new situations), analyze (breaking down concepts into constituent parts and verifying how they relate to each other), evaluate (making judgments using certain criteria), and create (making new products using previous learning experience). They are regarded as relevant skills for CT development that should be taught in EAP (de Chazal, 2014) and CLIL (Coyle et al., 2010). Among them, "analyze" and "evaluate" are most often associated with CT (de Chazal, 2014).

The ability to present arguments is an essential CT skill because it involves presenting one's views with both reasons and evidence (Chaffee, 2019; Fisher, 2011; Nosich, 2022). As Cottrell (2017) states, "essays are exercises in critical thinking" (p. 161). Notably, in argumentative essay writing tasks—with or without source materials—students must not only present their ideas but also assess their own reasoning. At a minimum, essay writing involves remembering (retrieving relevant information), creating (writing an essay), and evaluating (critiquing one's own ideas) (Anderson & Krathwohl, 2001). As fundamental CT abilities, critiquing one's own reasoning and engaging in dialectical thinking (Tanaka & Gilliland, 2017), as well as the need for refinement, make essay writing appropriate for assessing students' CT skills (Wade, 1995).

CT assessment criteria

Scholars have proposed various criteria for assessing CT skills, including cognitive thinking and reasoning skills. Chaffee (2019) and Fisher (2011) provided two criteria focusing on reasoning: whether the reasons support its conclusion (validity) and whether the reasons are true and acceptable (truth). An argument that includes accurate reasons that fully support the writer's claims is considered a sound argument. Furthermore, Paul and Elder (2014) proposed the following nine intellectual standards for assessing reasoning: (a) clarity of statements, (b) accuracy of information (i.e., truth), (c) precision of statements, (d) relevance of ideas, (e) depth of thoughts, (f) breadth of viewpoints, (g) logicalness (i.e., validity), (h) significance of information, and (i) fairness of arguments (see also Nosich, 2022). These were proposed for use by those who study CT to evaluate a given argument and improve the quality of their own reasoning. Thus, these criteria were not specifically designed for assessing the CT skills of L2 learners through their argumentative essays. Yanning (2017) developed a rating scale based on Paul and Elder's (2014) standards and implemented it to measure Chinese students' CT through their L2 argumentative essays. However, as the aim of the study was to gauge the effectiveness of a pedagogical approach, the appropriateness of the scale itself was not scrutinized.

Some scholars have proposed certain criteria to specifically assess CT skills through argumentative essays. Cottrell's (2017) description of critical writers enlists the following features of essays that reflect CT skills: presenting arguments clearly to make them comprehensible to readers, selecting the most controversial points to discuss in detail, placing arguments in logical order to emphasize the most controversial points, and using discourse markers to help readers understand the arguments. Additionally, the Washington State University (WSU) Center for Teaching, Learning, and Technology (2009) developed a rating scale for CT skills displayed in argumentative essays consisting of seven criteria with detailed descriptors. The rating scale examines students' (a) identification of an issue, (b) consideration of the issue's context, (c) presentation and assessment of supporting evidence, (d) integration of diverse perspectives, (e) presentation of their own perspectives, (f) identification of implications and consequences, and (g) communication of the message. These criteria were identified based on the practical experiences of WSU's staff members. This scale has also been included in writing-intensive courses in a U.S. university's general education curriculum (Morozov, 2011). Although Cottrell's (2017) description and WSU's rating scale connect CT skills with writing abilities, they were neither developed specifically to assess L2 learners' CT skills nor based on research. Hence, these criteria, developed for native English speakers, do not necessarily consider the characteristics of L2 students' writing.

Finally, Stapleton (2001) created a scheme to quantify CT as displayed in argumentative passages written in English by Japanese university students. This covers the key elements of CT and examines the numbers of (a) arguments presented (opinions and their reasons), (b) evidence given in support of each reason, (c) presentation of opposing arguments, (d) refutations of these counterarguments, and (e) any potential fallacies. Nevertheless, the quantified outcomes here do not necessarily reflect the essay's CT level or logical quality. For example, presenting numerous pieces of supporting evidence does not mean that the writer possesses high CT skills. Thus, the scheme cannot readily be adopted to measure L2 students' level of CT displayed in their argumentative essays.

In summary, a wide range of criteria has been suggested to assess CT skills based on theories conceptualizing CT. A significant limitation of the current CT criteria is that they are neither empirically derived nor supported for use in L2 educational contexts. Therefore, it remains unclear whether the suggested criteria are relevant to and salient in L2 essay writing assessments and whether other important criteria exist that have not yet been acknowledged.

Conceptualizing constructs for the development of a rating scale

Investigating raters' intuitive judgments of CT skills is helpful in forming its conceptualization in L2 educational contexts. An empirical investigation of raters' judgments would reveal the construct's components and facilitate the development of a rating scale. Researchers have identified the influential features of various constructs in applied linguistics, including oral fluency (e.g., Bosker et al., 2013), accentedness and comprehensibility (e.g., Saito et al., 2017), oral communicative ability (Sato, 2012; Sato & McNamara, 2019; McNamara, 1990), and writing proficiency (e.g., Cumming et al., 2001). These studies scrutinized raters' intuitive judgments of the targeted constructs without using descriptors and rigorous training to assess them.

Furthermore, they identified the influential components of raters' intuitive judgments of the constructs using one of the following three approaches. The first approach investigated the correlation between raters' judgments and objective measurements of the linguistic features of the performances (Bosker et al., 2013; Saito et al., 2017). The second approach examined the relationship between raters' judgments and their ratings of specific performance features (Sato, 2012; McNamara, 1990). The third approach required raters to judge performances and verbalize their rating process to identify features that affected their judgments (Sato & McNamara, 2019; Cumming et al., 2001). The first and second approaches identify features that unconsciously influence raters' judgments (e.g., McNamara, 1990). However, they do not consider the influence of other factors. The third approach compensates for this limitation. Nonetheless, analyzing verbal protocols may not be sufficient because raters' reports may not accurately represent the actual factors that affected their judgments (Gass & Mackey, 2017).

These studies have had important implications for the development of rating scales for oral fluency as well as overall speaking and writing proficiency. However, investigation of raters' judgments in the context of assessing L2 learners' CT through argumentative essay writing has not been conducted yet.

Theoretical background

Rating scales are tools, composed of criteria which assess test-takers' performance. Consequently, it is important to shortlist criteria that should be included in the scales by operationally defining the target construct and specifying its constituents. In general, a theory explicating the target construct is an important frame of reference for operationally defining it (Bachman & Palmer, 2010). However, there is no agreed theory of writing explaining the construct of L2 writing itself (Knoch, 2022), and none of the CT assessment models was developed specifically for L2 writing based on research (e.g., Paul & Elder, 2014). In this context, the empirically identified components of CT contribute to conceptualizing it for L2 writing assessment and can be included as features into a rating scale.

This study aimed to identify the features of argumentative essays that allow for the assessment of CT by investigating how readers judge the writers' CT skills. More specifically, the study addresses the following research questions (RQs):

1. What is the relationship between rater judgments of students' CT skills and their ratings using the assessment criteria for L2 writing proficiency?
2. What essay features do raters consider when judging students' CT skills?

The second and third approaches (see the "Conceptualizing Constructs for the Development of a Rating Scale" section) were applied to answer RQs 1 and 2, respectively. RQ1 is concerned with raters' intuitive judgments of writers' CT skills corresponding to any criteria used to measure L2 writing proficiency. Additionally, to explore any influential features other than the criteria uncovered by RQ1, a verbal protocol analysis was employed for RQ2. Therefore, this study aims to conceptualize CT by combining both research approaches and compensating for their limitations.

Methodology

The present study is exploratory research employing inductive reasoning, as its purpose is to identify the relevant and salient CT features of argumentative essays without applying existing CT theoretical frameworks. This study implemented a convergent mixed-methods research design, which involves the collection and analysis of both quantitative and qualitative data to merge the results (Creswell, 2015). More specifically, the scores for students' argumentative essays awarded by five raters were analyzed to examine the relationships among the assessment criteria. Additionally, the raters' verbal protocol data were analyzed to reveal the essay features that influenced the raters' judgments of the students' CT skills.

Participants

Students

Eighty-nine first- and second-year university students who attended CLIL courses participated in this study and took both pre- and post-test (see the "Data Collection Instrument" section). They were from two elementary, three lower-intermediate, and two upper-intermediate English courses. Based on their Test of English for Academic Purposes scores (a placement test), it was determined that the students' proficiency grades were roughly equivalent to levels A2 to B1 of the Common European Framework of Reference (CEFR) (Council of Europe, 2001). In the CLIL courses, the students were taught an academic subject selected by each instructor (e.g., Japanese culture and world Englishes) and encouraged to use the four English language skills (reading, listening, speaking, and writing). CT development is an explicit aim stated in the course syllabus, although the degree to which it was covered in the classes depended on each instructor. Moreover, details about CT—including its definition and assessment—were not included in the syllabus. The students signed an informed consent form to agree to have their essays used for research purposes.

Raters

Five native English speakers with work experience as examiners of the International English Language Testing System (IELTS) participated in the study as raters. Table 1 shows their background information. All raters had a Master's degree and at least 16 years' experience in English language teaching and 6 years' experience as IELTS examiners. They were chosen because this research examined the effectiveness of the CLIL program using the IELTS rating scale.

Data collection instrument

This study used students' performance data derived from a course evaluation project that examined the effectiveness of a CLIL program offered at a private Japanese

Table 1 Background information about the raters (N=5)

	R1	R2	R3	R4	R5
Age	41–50	31–40	41–50	41–50	51–60
Gender	Male	Male	Female	Male	Male
Nationality	British	British	Filipino	British	British
Academic degree	MA	MA	MA	MA	MA
Years of teaching experience	21–25	16–20	16–20	21–25	26 or above
Years of examiner experience	11–15	6–10	11–15	11–15	16–20

university. Students of seven 28-class (14-week) CLIL courses underwent identical speaking and writing tests on two occasions: during the 2nd/3rd class (pre-test) and the 27th class (post-test). This study then used the students' essays in the pre- and post-writing tests and their subsequent ratings in its analysis.

Pre- and post-tests to measure students' productive skills were developed for the course evaluation. A timed-independent writing task was developed by the researcher along with his colleague, in which students were instructed to write an essay to answer the following prompt: "What motivates students to study their subject at university? Give specific details and examples to explain your answer." They were asked to write approximately 300 words in 30 min using either a computer or pen and paper. This is a causal argumentative essay task in which students are required to speculate on the possible causes of a given phenomenon (Ramage et al., 2015). The task was considered suitable to elicit students' CT skills because it involved critiquing and refining one's reasoning while formulating arguments. Moreover, a similar writing task has also been employed by some well-known English proficiency tests (e.g., IELTS and the Test of English as a Foreign Language) to assess argumentation of L2 learners (Hirvela, 2017). The topic was selected because it was assumed that students do not need any specialized background knowledge to respond to it, but rather are able to use their creativity and personal examples to construct their argument.

Data collection procedure

Students were informed that the test's purpose was to examine improvements in their productive language skills following a one-semester CLIL course. They were also told that their test results would not affect their grades in this course. However, they were not informed that their CT skills would be assessed through the tests.

The handwritten essays were typed in Microsoft Word, and the same formatting style was applied to all of them, including those typed by the students themselves (Times New Roman, 12-point, single-spaced). Then, two essays each from the 70 students, who had produced, were collected, and hence a total of 140 essays (70 each from the pre- and post-tests) were procured. These students were chosen from among those who wrote more than 120 words, as it would have been difficult to assess multiple linguistic features and CT skills in shorter essays. Furthermore, the sample comprised approximately an equal number of students randomly selected from each of the three English courses (elementary: n = 23, lower-intermediate: n = 24, and upper-intermediate: n = 23). The essays ranged between 121 and 356 words, with an average word count of 205.6.

Next, the five raters were given the 140 essays for assessment. Each essay was scored by two to three of the five raters, with the connectivity required for a Rasch analysis being established. Each rater was requested to assess 60 essays. The five raters were asked to rate the essays using the IELTS Task 2 Writing band descriptors (public version) (British Council, n.d.), which consist of four criteria: Task Achievement, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy. The IELTS scale was used because it includes a wide range of writing proficiency components with detailed descriptors. Additionally, as Plakans and Gebril (2017) claim, argumentation could be measured by these criteria. Although the IELTS rating scale includes 10 levels

(0–9), this study only used six (1–6) because the students' proficiency was not high enough for them to obtain scores above seven.

In addition to assessing the essays according to the IELTS criteria, the raters were also asked to judge the level of each student's CT skills. The following two definitions of CT from online English dictionaries were provided:

- "The objective analysis and evaluation of an issue in order to form a judgement" (Oxford Dictionary).
- "The process of thinking carefully about a subject or idea, without allowing feelings or opinions to affect you" (Cambridge Dictionary).

These definitions include the key elements of CT addressed in L2 pedagogies, such as careful thinking (reflective thinking), objective analysis and evaluation (rational thinking), and judgment formation (reasonable thinking) (Dummett & Hughes, 2019). Definitions from English dictionaries, rather than those found in the CT literature, were provided because they are concise and easy to understand. The raters scored the students' CT skills using a six-level semantic differential scale ranging from one "Poor" to six "Excellent" with unspecified midpoints. Descriptors and training in using the scale were not provided because this study's main aim was to investigate raters' interpretation of students' CT without the influence of any pre-existing rating scale. The raters were informed that this criterion was adopted for research purposes rather than for course evaluation.

Each rater performed a retrospective written verbal protocol (Gass & Mackey, 2017) by writing comments on eight to 10 randomly selected essays. The raters were asked to indicate which part of the students' essays influenced the judgments of their CT skills and explain how these identified portions influenced the scores assigned using Microsoft Word's comment function immediately after scoring the essays. In total, 103 comments were given to 18 essays. Unfortunately, a written verbal protocol could undermine the validity of the reports because the raters would be able to write about things that they did not think about while actually rating students' CT skills. However, an oral concurrent verbal protocol could not be adopted because the raters concurrently scored other features of the essays, whereas the focus of this study was only on CT.

Data analysis

The scores given to the 140 essays were statistically analyzed to answer RQ1. First, the rater reliability was confirmed using the many-facet Rasch measurement. The pre- and post-test data were separately analyzed using the FACETS 3.83.0 software (Linacre, 2019). Rater infit mean-square values, which indicate rater reliability, were within the acceptable range (0.7–1.3). This suggests that all the raters scored the students' argumentative essays consistently using the IELTS band descriptors and a scale for assessing CT skills. Second, a multiple regression (MR) was conducted using the raw scores to examine the relative importance of the four IELTS criteria (the predictor variables) in predicting the raters' judgments of students' CT skills (the outcome variable). As there is no hypothesis about the strength of the predictors, this study performed a standard

MR simultaneously with all predictors. The pre- and post-test data were separately analyzed using SPSS Statistics version 26. The assumptions for the MR (the number of data cases, multicollinearity, normality, linearity, and homoscedasticity) were examined. This study had 150 data points each in the pre- and post-tests, which was larger than the required 15 cases of data per predictor (Field, 2018). As the variance inflation factor values ranged from 1.89 to 2.74, staying far under 10, multicollinearity was not present among the predictor variables. The last three assumptions were examined using the scatterplots of residual, histogram, and P-P plot. All the assumptions were satisfied except for normally distributed errors for both the pre- and post-tests. However, the violation of this assumption is not of great concern because of the amount of data in this study (Field, 2018).

The raters' verbal protocol data were then analyzed to answer RQ2. Thematic analysis, which involves identifying themes within the data (Braun & Clarke, 2013), was carried out to identify the features of students' argumentative essays affecting their CT scores as assessed by the raters. First, each of the raters' comments was read to generate initial codes that grouped similar concepts. Second, coding categories, based on the generated codes, were developed, with all comments being sorted into the developed categories using NVivo 11. Third, the coding categories were reexamined and collated to identify any overarching themes. Therefore, the analysis was inductive, with the identified themes being linked to the data. Finally, a PhD student in applied linguistics was asked to code 30% of the data to ensure inter-coder reliability. The kappa coefficient was 0.71, demonstrating adequate agreement. Disagreements were resolved through subsequent discussions, and the categories were finalized.

Results

This section will present the results of the MR answering RQ1 (identifying the relationship between rater judgments of students' CT skills and their ratings of other criteria for L2 writing proficiency) and the thematic analysis answering RQ2 (exploring the essay features that the raters consider while judging students' CT skills).

Relationship between CT and other criteria

Tables 2 and 3 present the descriptive statistics of the scores on each criterion and the results of the MR analyses, respectively. The regression results indicate that the Task Achievement scores made the largest contribution to the raters' intuitive

Table 2 Means and standard deviations of scores for each criterion

Criteria	Pre-test		Post-test	
	M	*SD*	*M*	*SD*
Critical thinking skills	3.49	0.69	3.54	0.67
Task achievement	4.89	0.83	5.07	0.70
Coherence & cohesion	4.81	0.64	4.97	0.60
Lexical resource	4.75	0.56	4.87	0.57
Grammatical range & accuracy	4.74	0.57	4.90	0.56

Note. N of cases = 150

Table 3 Standard MR results

Variable	*B*	95% CI for *B*		*SE B*	*β*	*p*	R^2
		LL	*UL*				
Pre-test							.47
Constant	−0.25	−1.03	0.53	0.40		.524	
Achievement	0.32	0.17	0.48	0.08	.39	.000	
Coherence	0.14	−0.05	0.33	0.10	.13	.135	
Lexis	0.18	−0.06	0.42	0.12	.15	.142	
Grammar	0.13	−0.08	0.34	0.10	.11	.213	
Post-test							.33
Constant	−0.04	−0.95	0.87	0.46		.935	
Achievement	0.29	0.11	0.46	0.09	.30	.002	
Coherence	0.07	−0.16	0.29	0.11	.06	.551	
Lexis	0.34	0.10	0.59	0.12	.29	.006	
Grammar	0.03	−0.21	0.26	0.12	.02	.821	

Note. *CI* confidence interval, *LL* lower limit, *UL* upper limit, *Achievement* Task Achievement, *Coherence* Coherence and Cohesion, *Lexis* Lexical Resource, *Grammar* Grammatical Range and Accuracy

judgments of CT skills in both the pre- and post-tests (*β*s = .39 and .30, *p*s = .000 and .002, respectively). Additionally, the contribution of Lexical Resource was statistically significant and found to be the second largest predictor in the post-test (*β* = .29, *p* = .006). In contrast, the other criteria had minor contributions and were not significant (*p*s > .05). Overall, the four IELTS criteria explained 47% and 33% of the variance in CT skills in the pre- and post-tests, respectively, suggesting that elements other than the IELTS criteria explain rater judgments of CT. To summarize, the raters' judgments of participants' CT skills were explained by their Task Achievement scores most strongly, followed by the Lexical Resources scores.

Table 4 Categories of influential features of argumentative essays

Category	Definition	Example comments
Relevancy	Connection between ideas and the theme	"The student attempts to make a valid point that is related to the topic." (R3) "Not really related to the question." (R1)
Content development	The amount of ideas, supports, examples, etc. and the depth of ideas	"This idea is quite well developed and argued." (R1) "This is a very short paragraph that lacks development and therefore lacks critical thinking." (R2)
Logicality	Logical structure, connection between ideas, and reasons supporting claims	"Some CT: evidence of logical reasoning." (R5) "These sentences are poorly linked together showing no obvious connection." (R2)
Quality of idea	Validity of ideas, originality of ideas, and the range of perspectives	"Almost childlike thinking here!" (R1) "These ideas are common and universal, thus the average rating of 3." (R4) "This student is thinking only in terms of their own experience and not the wider concepts." (R1)
Other features	Linguistic accuracy and miscellaneous features	"Again, awkward vocabulary of "hamsam" (handsome?)." (R2) "There is some variation in the comment about unmotivated students." (R4)

Influential features on rater judgments of CT

Table 4 shows the findings of the thematic analysis of the comments written by the raters. The following five features, representing the criteria used by the raters to judge the students' CT skills, were explored: Relevancy, Content Development, Logicality, Quality of Ideas, and Other Features. The first category, Relevancy, concerned the question of whether the written ideas were addressing the given question (What motivates students to study their subject at the university?). The raters positively evaluated essays that maintain their focus on the question and negatively judged pieces of content deviated from it. Second, Content Development referred to how deeply students discussed their ideas by including supporting details and examples. Essays with a sufficient amount of details, examples, and ideas were considered as those displaying high CT skills. Third, the raters noted the logicality of the arguments, the link between written ideas, and coherence. The raters considered that high CT skills were demonstrated by logical connections among ideas, especially the link between the writers' main claim and supporting evidence. Fourth, the raters evaluated the quality of ideas focusing on the validity and originality of ideas as well as on the width of perspective presented in the essays. Students discussing well-thought and original ideas from multiple points of view were regarded as those possessing high CT skills, whereas those presenting poor and ubiquitous thoughts based only on their own personal experience were evaluated otherwise. Finally, linguistic accuracy and miscellaneous features were categorized as Other Features.

Three raters (R1, R2, and R3) made some comments on linguistic errors found in the essays. For example, R3 pointed out linguistic accuracy by saying, "Despite the inaccuracies in language and grammar, the student is able to present a weak link between motivation and being able to pursue one's own interests." However, linguistic features were not regarded as an independent factor that influences rater judgments of students' CT skills. First, comments on linguistic errors were not prevalent within the protocol data (5.2% of all the comments). Second, half of the comments on linguistic errors were in *although* clauses or *despite* phrases as in the example above, suggesting that the influence of linguistic features may be weaker than that of the other features presented in Table 4.

To illustrate the essay features that affected raters' judgments of students' CT skills, three essays, and the corresponding raters' feedback, are presented in Tables 5, 6, and

Table 5 Body paragraph of the essay written by Student 15 (left) and raters' comments (right)

First, I explain curious. If we don't like math, we don't have curious for math. At that time, if we study math, we don't keep concentration, and don't feel interesting. I think this lead to down the motivation for studying. From this, curious is very important for studying. Second, I explain feelings. It also lead to curious, if we feel the studying is interesting, we want to study more and more. Moreover, if we study feeling like this, the academic world spread and studying is more interesting. For example, if we learn the literature of Chinese which written a long time ago, we can the way Chinese think a long time ago, and we may be interesting in the changes the way Chinese think throughout time. Like this, we can feel interesting that study lead to study, it lead to motivation.	1. Good: exploring and exemplifying an idea. (R1) 2. some CT: logical reasoning (R5) 3. The student tries to demonstrate the connection between 'curious' and 'motivation for studying'. He/She tries to give structured support for this particular issue. (R3) 4. The student tries to illustrate the connection between 'feelings' and 'motivation'. He/she attempts to identify the main issues related to this particular argument and then makes a reasonable attempt to link these different factors. (R3) 5. Same again: although there are some issues with clarity here. (R1) 6. some CT: example used as support is relevant and shows original thinking (R5)

Table 6 The essay written by Student 69 (left) and raters' comments (right)

I think dream motivates students to study their subject at university. I major "Material Life and Science". I study Chemistry, Physics, and biology. Studying Science is very hard for me because it is difficult. But I want to be a scientist in the future. So, I can do my best to study science. And we often do experiment about chemistry, Physics, and biorogy. It is very fun. And I think GPA motivates student to study their subject at university. If we couldn't get good GPA, we can't enter the room of experiment which we want to enter. So, we have to study hard about subject which we are not interested in. If I could not enter the room which we are not interesting in, I'm very sad and I can't do my best. So, I study hard now to become a sicentist.	1. The student attempts to make a valid point that is related to the topic. (R3) 2. low CT: explanation of context egocentric (R5) 3. It is unfortunate that the student chooses to focus entirely on his/her own experiences and circumstances. Although these are not wholly unrelated, the student could have extended his/her argument by including general issues which are related to the main topic. (R3) 4. The 10th sentence (about GPA) could have been given more logical support resulting in what would have been a more coherent essay. (R3) 5. some CT: considers implications and impact on other people (R5) 6. It's not that CT skills are poor, rather there is not much on display. The ideas are kind of common sense, only developed very simply and entirely from the writer's own experience. (R1)

7. These essays received positive, positive and negative, and negative comments, respectively. Additionally, they included a wide range of features, as presented in Table 4.

Table 5 presents the body paragraph of the essay written by Student 15 in the pretest. The essay was rated by R1, R3, and R5, and the scores for CT skills given by the raters were 5, 4, and 3.5, respectively. The body paragraph contained two factors that motivate students to study their subjects at the university and supporting details. Comments 1 to 3 were given to the first factor, while Comments 4 to 6 were given to the second factor.

In the essay, Student 15 argues that curiosity motivates students to study at the university and explains it by providing a negative case in which students are not curious about math, which eventually leads to less concentration, interest, and motivation. In the latter half of the paragraph, she points out that positive feelings toward learning motivate students to study and presents a concrete example of how learning leads to more interest in the subjects. Overall, the raters' comments on Student 15's CT skills were positive. Raters appeared to perceive that both factors were supported by logical reasoning and relevant examples, which positively contributed to their judgments of her CT skills. For example, a chain of reasoning explaining why curiosity is important (from second to fifth sentences) was perceived as logical and connected to motivation for studying. The second argument (positive feelings toward learning) was also judged to be connected to motivation

Table 7 Body paragraphs of the essay written by Student 50 (left) and raters' comments (right)

First, it is future dream. I have a dream. I want to be a botanist. But it have the problem that I am not good at speaking and writing English. For, botanist should read many book that are read English and speak my researches in English. Second, it is owe to future to enjoy. I went to Australia three years ago. However, I could not speak English well. So I want to revenge. And I want to go abroad because of studying science. Foreign countries have many animals and nature. I have ever seen foreign countries nature because I still stayed home in Australia. So I want to a lot of nature. Third, it is what I make my friends. I have a few friends. So I want to make friends of alien. And I want to speak English with. And I want to discuss science in English.	1. These opinions bear no relation to the question, so it's hard to rate the response for critical thinking. The writer has not appropriately engaged with the topic. (R2) 2. Again, these ideas are basic and universal, thus the average rating of 3. (R4)

by R3 and R6. Furthermore, Comment 6 made by R5 indicates that the originality of the idea was part of CT from the raters' perspective. The example of learning about Chinese literature was considered original and evaluated positively. Simply presenting ubiquitous arguments and supporting details may give the impression that students did not consider the given question carefully.

Table 6 presents the entire essay written by Student 69 in the post-test. The essay was rated by R1, R3, and R5, and the scores for CT skills given by the raters were 3, 3.5, and 3, respectively. It contained two factors that motivate students to study at the university and supporting details. Comments 1 to 3 were given to the first factor, while Comments 4 to 6 were given to the second factor.

In the essay, Student 69 argues that students' dreams and Grade Point Averages (GPAs) motivate them to study at the university. First, she claims that university students' dreams motivate them to study by providing a personal example in which she is able to study science hard because being a scientist is her dream for the future. Second, the student mentions that GPA is an incentive to study as students cannot study at the laboratory they wish if they have a low GPA. Raters acknowledged that her arguments successfully addressed the question (Comments 1 and 3). However, they negatively commented that the supporting evidence was based primarily on the student's personal experience, therefore considering it egocentric. Although the support for the second factor was positively judged by R5 (as the student explains how low GPA influences all university students and not only herself), R3 commented that the argument should have been supported with more logical reasoning. The inclusion of her personal feeling ("If I could not enter the room which we are not interesting in, I'm very sad and I can't do my best.") may have made the second factor sound less logical and coherent. Finally, R1 wrote that the essay does not display the student's CT skills (Comment 6). In a different essay, he also noted: "Perhaps it is difficult to show great CT skills with this task, as they are not really analyzing a text or doing any research." This suggests that R1 appears to believe that timed independent essay writing cannot appropriately elicit the writer's CT skills.

Table 7 presents the body paragraphs in the essay written by Student 50 in the pretest. The essay was rated by R2 and R4, and the scores for CT skills given by the raters were 2 and 3, respectively. The comments refer to the entire essay.

In his essay, Student 50 discusses three points of personal dream plan for future, plan to travel abroad, and desire to make friends. However, he fails to explain clearly and explicitly how the three points motivate students to study. R2 perceived that these points were not relevant to the question and evaluated that the student did not engage in the topic appropriately. R2 also mentioned that it was difficult to rate the student's CT skills because of the irrelevant opinions presented. Although his arguments are based solely on his personal experience as in Student 69's essay, this feature was not mentioned by the raters. Furthermore, R4 commented that the three points raised by the student are basic and universal, which influenced his rating of the student's CT skills. As discussed above, the presentation of universal opinions may negatively affect the raters' impression of the writer's CT skills. Nevertheless, it was not clear how the raters judged the extent to which written thoughts were universal or original.

Discussion

RQ1 asked: "What is the relationship between rater judgments of students' CT skills and their ratings on the criteria used to measure L2 writing proficiency?" The results indicate that raters' judgments of students' CT skills are most strongly explained by Task Achievement scores, although Lexical Resource scores were found to be another significant predictor in the post-test.

Task Achievement measures how adequately a student addresses all parts of the task, presents their position, and develops their main ideas with relevant details (British Council, n.d.). Therefore, in the argumentative essay task used in this study, this criterion concerned itself with the extent to which students adequately answered the prompt and supported their answers by giving relevant and specific details, as well as examples. As some literature indicates, these elements are related to CT. Specifically, these are equivalent to two intellectual standards proposed by Paul and Elder (2014): relevance (how well the idea is connected to the question) and clarity (how well the idea is explained and elaborated). Moreover, Task Achievement appears to involve some aspects recognized in Stapleton's (2001) scheme: the presence of arguments (opinions and their reasons) and supporting evidence. This finding suggests that the raters' judgment of writers' CT skills might be influenced by the content of argument more than how it is presented even in learners' essays including linguistic errors.

This study found a weak relationship between CT skills and the linguistic features displayed in the participants' essays, suggesting that demonstrating a high linguistic quality does not guarantee positive judgments of CT skills from readers. This supports the claim made by de Chazal (2014) that language proficiency is not a predictor of CT ability. However, using diverse and accurate vocabulary, measured by the Lexical Resource criterion, may lead to better impressions of one's CT skills on readers. This may be since diverse vocabulary results in development of an idea, which was judged as a relevant element of CT. Additionally, errors in vocabulary in the L2 students' essays might have undermined the clarity and comprehensibility of the content. As the clarity of statement is a fundamental element in the sense that other elements cannot be evaluated unless the content is written clearly (Nosich, 2022; Paul & Elder, 2014), the use of vocabulary influencing accurate conveyance of messages could be a linguistic feature relevant to CT skills especially in L2 argumentative writing. However, it remains unknown why Lexical Resource scores were not a significant predictor of CT in the pre-test.

To further examine the essay features that contributed to raters' judgments of CT, RQ2 asked, "What essay features do raters consider when judging students' CT skills?" The analysis of the protocol data revealed five categories: relevancy to the question, content development, logicality, quality of ideas, and other features. The first two categories support the results of RQ1 and align with the concepts of relevancy and clarity in Paul and Elder's (2014) criteria. Few comments on linguistic features also partially concur with the outcome of RQ1. Overall, the raters seemed to construe CT skills displayed through the writing task as argumentation skills as emphasized in critical thinking movement (Davies & Barnett, 2015) and the CT literature (e.g., Cottrell, 2017; Fisher, 2011; Nosich, 2022).

Rater judgments of CT skills also included elements that were not addressed by the Task Achievement criterion: logicality and quality of ideas (see Table 4). Comments on logicality (logical structure, connection between ideas, and reasons supporting claims) showed that raters seemed to focus on the logical reasoning supporting students' claims and fallacies, which is regarded as an assessment criterion for both arguments and CT skills (Paul & Elder, 2014; Stapleton, 2001). This is partly addressed by the Coherence and Cohesion criterion in the IELTS rating scale (logical sequencing of information and ideas) (British Council, n.d.). This feature was considered as an essential component of CT in the argumentative writing task, in which "an author states a claim, uses some form of evidence—data, reasons, examples, etc.—to support the claim, and shows how the evidence supports the claim" (Hirvela & Belcher, 2021, p. 1). The central purpose of the writing task could influence the raters' attention to logicality. Moreover, the raters might have applied their critical reading skills, which involve appraising the degree to which the students adequately justified their opinions (Wallace & Wray, 2021).

The quality of ideas was primarily related to the range of perspectives displayed in the essay and their originality. First, supporting a claim by simply citing personal experience was judged negatively and regarded as egocentric (see Table 6). In contrast, raters positively evaluated writers who explained how a certain factor motivates university students in general, not solely for them, to study at university. This suggests that the type of evidence used influences rater judgments of a writer's CT, and anecdotal evidence can be perceived less persuasive than other types, including causal evidence (Hornikx & Hoeken, 2007). Additionally, supporting claims via personal experiences can be perceived as failing to consider the question from other perspectives. This may negatively influence the judgment of a writer's CT skills, as engaging in broader thinking by seeing situations from different perspectives has been identified as a key component of CT (Chaffee, 2019; Nosich, 2022; Paul & Elder, 2014). Second, raters positively judged original thoughts but negatively evaluated common and universal ideas. This is related to a disposition of critical thinkers known as intellectual autonomy (Paul & Elder, 2014), which entails having authorship of one's own thoughts rather than simply accepting or borrowing those of others. Raters' focus on originality of thought resonates with a conceptualization of CT given by academics of history, philosophy, and literary/cultural studies in Moore's (2013) study. In particular, originality may be relevant to tasks involving creating or producing ideas (Anderson & Krathwohl, 2001), including the argumentative writing used in this study.

It was found that raters neither mentioned all of the features recognized in the literature nor focused on the same features in the essays written by different students. For example, the raters seldom commented on the accuracy of supporting evidence (Paul & Elder, 2014) and did not comment on the breadth of thinking or the inclusion of counterarguments in the essays (Stapleton, 2001; Washington State University Center for Teaching, Learning, and Technology, 2009). This may be because only a few students supported their ideas using evidence other than personal experiences and included counterarguments that challenged their own points. It suggests that relevant and salient CT criteria depend on writing tasks (e.g., independent or integrated). In this study, the raters' foci might be narrower than theoretical models because the students were required to write an argument, without any external resources, in a short period of time. Furthermore, raters focused on the essay's relevancy to the question in those written by students who obtained low CT scores (Table 7) and focused more on logicality and the quality of ideas in essays written by those who obtained medium to high CT scores

(Table 5). This suggests that the essay features influencing raters' judgments depend, not only on CT features displayed in the essays, but also on the students' overall CT level.

Conclusion

This study investigated five raters' judgments of students' CT skills through reading and rating their argumentative essays, thereby revealing the features of the essays that contributed to their judgments. The results indicate that the raters' intuitive perceptions of students' CT skills were linked to the Task Achievement and, partly, Lexical Resource criteria in the IELTS (British Council, n.d.). Additionally, raters' written comments revealed that their judgments of the writers' CT skills were affected by the essay's relevancy to the question, content development, logicality, and quality of ideas.

The findings of this study help to delineate the CT skills addressed in L2 pedagogies so that they can be assessed through argumentative essays. In particular, test developers and teachers who are keen to assess test-takers' CT skills could incorporate the explored features into the assessment criteria. Since there are distinct elements pertinent to CT skills, it is possible to address them in different criteria for assessing essays. If a pre-established rating scale needs to be adopted because of practical constraints, the Task Achievement category in the IELTS rating scale (British Council, n.d.), addressing relevancy and content development, can be a viable option as scores predict students' CT skills to some degree. However, it is recommended to incorporate other features explored by this study into rating scales for a more precise measurement of CT skills, because positive judgments from raters are likely to require more than what the Task Achievement category comprehends, such as logicality, range of perspectives (or types of evidence), and originality of ideas. Logicality can be assessed through the category of coherence, which typically focuses on progression of ideas and logical sequencing (Knoch, 2007). Although it already entails some components of logicality explored in the study, the category can explicitly mention the connection between a claim and supporting evidence to measure CT more precisely. The quality of ideas and originality could be evaluated though the category of content. While some existing content categories address relevancy and content development (e.g., Jacobs et al., 1981), the creativity of ideas and their ability to create interest can also be addressed in the criterion (see Bae et al., 2016), although the assessment of these features is highly subjective.

It is also possible to create a single-independent assessment criterion for CT by including all the features discovered by the study. This option is beneficial because the same features are not necessarily relevant and salient across different proficiency levels. This study indicated that raters focused on the essay's relevancy to the question written by students who obtained low CT scores and focused more on logicality and quality of ideas in the essays written by those who obtained medium to high CT scores. This finding suggests that descriptors for low CT levels should focus on the essay's relevancy to the topic, while descriptors for higher levels should address the logicality and originality in the content.

This study's findings also have some implications for classroom-based assessment activities, especially for self- and peer-assessment (Brown & Abeywickrama, 2019). In the EAP and CLIL classrooms, which emphasize cultivating CT skills, self- and peer-assessment is recognized as a useful activity for improving students' task performance (e.g., Coyle et al., 2010; Ferris, 2018). The explored features can be incorporated into the development of checklists, which convey the students the construct in a simple manner and are suitable for use in self- and peer-assessment (Green, 2021). For instance, after reading their own or peers' argumentative essays, students could be asked to respond to statements such as "The essay discusses the given prompt without any irrelevant piece of content," or "The essay includes opinions from various perspectives" by indicating yes or no. Can-do statements can also be created based on the study's findings for self-assessment, including statements such as "I can give sufficient examples supporting my opinion" or "I can connect ideas logically." In so doing, students can grasp the elusive concept of CT and realize the characteristics of highly evaluated essays without deeply learning about the definition of CT itself.

This study has some limitations. First, this study collected data from only five native English speakers who have worked as IELTS examiners. Because of this small sample size, it is difficult to generalize these findings in broader contexts, including readers with different backgrounds. Moreover, the raters did not necessarily have a deep understanding of CT and exhibited raters' bias in their ratings. Second, this study collected data from students with relatively low English proficiency levels. Students with higher English proficiency may display a wider range of features related to CT skills, including reflecting on their own supporting evidence (Chason et al., 2017). Third, written verbal protocol data may suffer from non-veridicality, such that the results reported may have included features that raters did not actually consider while rating or may not have comprehensively included all of the features that they considered (Gass & Mackey, 2017). Post-marking interviews should have been conducted to triangulate the findings. Fourth, the assigned essay topic and writing task also had some limitations. Whether or not the students were able to demonstrate their CT skills through the topic was not determined. In other words, the topic's appropriateness was not examined. Lastly, only one type of writing assignment (timed, independent, causal argumentative essay writing) was employed to examine the relevant criteria for measuring CT skills. As noted by R1, the task adopted in this study did not involve any analysis and research, meaning that it may not always be suitable to assess CT. Using other types of essays, such as source-based argumentative writing (Plakans & Ohta, 2021), might reveal different dimensions of CT, that is, different essay features might have been found to be relevant to raters' judgments of CT skills.

Therefore, further research is recommended to investigate the way in which CT skills are related to various essay writing tasks, including integrated writing tasks or research projects, which can be done by examining rater judgments and collecting data that compensate for limitations of verbal protocols (e.g., interviews). Such research will help in revealing the elusive concept of CT skills in L2 pedagogies.

Abbreviations

CEFR: Common European Framework of Reference; CLIL: Content and Language Integrated Learning; CT: Critical thinking; EAP: English for academic purposes; GPA: Grade point average; IELTS: International English Language Testing System; L2: Second language; MR: Multiple regression; RQ: Research question; WSU: Washington State University

Acknowledgements

Not applicable.

Author's contributions

The author read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

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

Received: 5 January 2022 Accepted: 4 March 2022
Published online: 08 April 2022

References

Anderson, L., & Krathwohl, D. (Eds.) (2001). *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives*. London: Pearson.
Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.
Bae, J., Bentler, P. M., & Lee, Y.-S. (2016). On the role of content in writing assessment. *Language Assessment Quarterly*, *13*(4), 302–328. https://doi.org/10.1080/15434303.2016.1246552.
Bosker, H. R., Pinget, A., Quené, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*(2), 159–175. https://doi.org/10.1177/0265532212455394.
Braun, V., & Clarke, V. (2013). *Successful qualitative research: a practical guide for beginners*. London: Sage.
British Council. (n.d.). *IELTS task 2 writing band descriptors (public version)*. Retrieved October 1, 2018, from https://takeielts.britishcouncil.org/sites/default/files/ielts_task_2_writing_band_descriptors.pdf#search=%27IELTS+TASK+2+Writing+band+descriptors+%28public+version%29%27.
Brown, H. D., & Abeywickrama, P. (2019). *Language assessment: principles and classroom practices*, (3rd ed.). Hoboken: Pearson.
Chaffee, J. (2019). *Thinking critically*, (12th ed.). Boston: Cengage Learning.
Chason, L., Loyet, D., Sorenson, L., & Stoops, A. (2017). An approach for embedding critical thinking in second language paragraph writing. *TESOL Journal*, *8*(3), 582–612. https://doi.org/10.1002/tesj.288.
Cottrell, S. (2017). *Critical thinking skills: effective analysis, argument and reflection*, (3rd ed.). London: Red Globe Press.
Council of Europe (2001). *Common European Framework of Reference for languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.
Coyle, D., Hood, P., & Marsh, D. (2010). *CLIL: Content and language integrated learning*. Cambridge: Cambridge University Press.
Creswell, J. W. (2015). *A concise introduction to mixed methods research*. Thousand Oaks: Sage.
Cumming, A. H., Kantor, R., & Powers, D. E. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: an investigation into raters' decision making and development of a preliminary analytic framework. (TOEFL Monograph No. MS-22)*. Princeton: Educational Testing Service.
Davies, M., & Barnett, R. (2015). Introduction. In M. Davies, & R. Barnett (Eds.), *The Palgrave handbook of critical thinking in higher education*, (pp. 1–25). New York: Palgrave Macmillan.
de Chazal, E. (2014). *English for academic purposes*. Oxford: Oxford University Press.
Dummett, P., & Hughes, J. (2019). *Critical thinking in ELT: a working model for the classroom*. Boston: National Geographic Learning.
Ennis, R. (2011). Critical thinking: reflection and perspective Part I. *Inquiry: Critical Thinking Across the Disciplines*, *26*(1), 4–18. https://doi.org/10.5840/inquiryctnews20112613.
Ferris, D. R. (2018). Writing instruction and assessment: activities, feedback, and options. In J. M. Newton, D. R. Ferris, C. C. M. Goh, W. Grabe, F. L. Stoller, & L. Vandergrift (Eds.), *Teaching English to second language learners in academic contexts: reading, writing, listening, and speaking*, (pp. 106–122). New York: Routledge.
Field, A. (2018). *Discovering statistics using IBM SPSS statistics*, (5th ed.). London: Sage.
Fisher, A. (2011). *Critical thinking: an introduction*, (2nd ed.). Cambridge: Cambridge University Press.
Gass, S. M., & Mackey, A. (2017). *Stimulated recall methodology in applied linguistics and L2 research*, (2nd ed.). New York: Routledge.
Green, A. (2021). *Exploring language assessment and testing: language in action*, (2nd ed.). Oxon: Routledge.
Hirvela, A. (2017). Argumentation and second language writing: are we missing the boat? *Journal of Second Language Writing*, *36*, 69–74. https://doi.org/10.1016/j.jslw.2017.05.002.
Hirvela, A., & Belcher, D. (2021). Introduction. In A. Hirvela, & D. Belcher (Eds.), *Argumentative writing in a second language: perspectives on research and pedagogy*, (pp. 1–9). Ann Arbor: University of Michigan Press.
Hornikx, J., & Hoeken, H. (2007). Cultural differences in the persuasiveness of evidence types and evidence quality. *Communication Monographs*, *74*(4), 443–463. https://doi.org/10.1080/03637750701716578.
Jacobs, H., Zinkgraf, S., Wormuth, D., Hartfiel, V., & Hughey, J. (1981). *Testing ESL composition: a practical approach*. Rowley: Newbury House.
Knoch, U. (2007). 'Little coherence, considerable strain for reader': a comparison between two rating scales for the assessment of coherence. *Assessing Writing*, *12*(2), 108–128. https://doi.org/10.1016/j.asw.2007.07.002.
Knoch, U. (2022). Assessing writing. In G. Fulcher, & L. Harding (Eds.), *The Routledge handbook of language testing*, (2nd ed., pp. 236–253). Oxon: Routledge.
Linacre, J. M. (2019). *Facets computer program for many-facet Rasch measurement, version 3.83.0 [Computer Software]*. Winsteps.com.
McNamara, T. F. (1990). Item Response Theory and the validation of an ESP test for health professionals. *Language Testing*, *7*(1), 52–75. https://doi.org/10.1177/026553229000700105.

McNamara, T. F. (1996). *Measuring second language performance*. London: Pearson.

Mehisto, P., & Ting, T. (2017). *CLIL essentials for secondary school teachers*. Cambridge: Cambridge University Press.

Moore, T. (2013). Critical thinking: seven definitions in search of a concept. *Studies in Higher Education*, *38*(4), 506–522. https://doi.org/10.1080/03075079.2011.586995.

Morozov, A. (2011). Student attitudes toward the assessment criteria in writing-intensive college courses. *Assessing Writing*, *16*(1), 6–31. https://doi.org/10.1016/j.asw.2010.09.001.

Nosich, G. (2022). *Critical writing: a guide to writing a paper using the concepts and processes of critical thinking*. London: Rowman and Littlefield.

Paul, R., & Elder, L. (2014). *Critical thinking: tools for taking charge of your learning and your life*, (2nd ed.). Lanham: Rowman and Littlefield.

Plakans, L., & Gebril, A. (2017). An assessment perspective on argumentation in writing. *Journal of Second Language Writing*, *36*, 85–86. https://doi.org/10.1016/j.jslw.2017.05.008.

Plakans, L., & Ohta, R. (2021). Source-based argumentative writing assessment. In A. Hirvela, & D. Belcher (Eds.), *Argumentative writing in a second language: perspectives on research and pedagogy*, (pp. 64–81). Ann Arbor: University of Michigan Press.

Ramage, J. D., Bean, J. C., & Johnson, J. (2015). *Writing arguments: a rhetoric with readings*, (10th ed.). Boston: Pearson.

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: a validation and generalization study. *Applied Linguistics*, *38*(4), 439–462. https://doi.org/10.1093/applin/amv047.

Sato, T. (2012). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing*, *29*(2), 223–241. https://doi.org/10.1177/0265532211421162.

Sato, T., & McNamara, T. (2019). What counts in second language oral communication ability? The perspective of linguistic laypersons. *Applied Linguistics*, *40*(6), 894–916. https://doi.org/10.1093/applin/amy032.

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.

Tanaka, J., & Gilliland, B. (2017). Critical thinking instruction in English for academic purposes writing courses: a dialectical thinking approach. *TESOL Journal*, *8*(3), 657–674. https://doi.org/10.1002/tesj.291.

Wade, C. (1995). Using writing to develop and assess critical thinking. *Teaching of Psychology*, *22*(1), 24–28. https://doi.org/10.1207/s15328023top2201_8.

Wallace, M., & Wray, A. (2021). *Critical reading and writing for postgraduates*, (4th ed.). London: Sage Publication.

Washington State University Center for Teaching, Learning, and Technology. (2009). *Critical and integrative thinking*. Retrieved March 3, 2020, from https://www.colorado.edu/sei/sites/default/files/attached-files/wsu-critical-thinking-rubric-2009.pdf.

Wilson, K. (2016). Critical reading, critical thinking: delicate scaffolding in English for Academic Purposes (EAP). *Thinking Skills and Creativity*, *16*, 256–265. https://doi.org/10.1016/j.tsc.2016.10.002.

Yanning, D. (2017). Teaching and assessing critical thinking in second language writing: an infusion approach. *Chinese Journal of Applied Linguistics*, *40*(4), 431–451. https://doi.org/10.1515/cjal-2017-0025.

Yuan, R., & Stapleton, P. (2020). Student teachers' perceptions of critical thinking and its teaching. *ELT Journal*, *74*(1), 40–48. https://doi.org/10.1093/elt/ccz044.

Publisher's Note

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