International Journal of Linguistics, Literature and Translation
ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print)
DOI: 10.32996/ijllt
Journal Homepage: www.al-kindipublisher.com/index.php/ijllt

IJLLT
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Second Language Students' Critical Thinking Performance in Argumentative Writing

Sheng Tan
PhD graduate, Faculty of Education, The University of Hong Kong, Hong Kong, China
Corresponding Author: Sheng Tan, **E-mail**: u3004292@connect.hku.hk

| ABSTRACT
Critical thinking (CT) is interconnected with argumentation, so argumentative writing serves as a crucial medium for demonstrating CT. The present study aimed to examine second language (L2) writers' CT performance in argumentative writing and to explore how high-achieving L2 students differ from their low-achieving counterparts in terms of CT performance. In this study, a sample of 33 higher-level and 32 lower-level L2 students from a Chinese university wrote an L2 argumentative essay on computers within 40 minutes. The CT performance in four major parts of each essay, i.e., position(s), explanations, evidence, and conclusion, was assessed according to four CT criteria: Unambiguity, Fair-mindedness, Substance, and Consistency. Findings suggest that (1) the performance of evidence and Substance was not satisfactory, (2) no significant differences were reflected in the performance of overall L2 CT, Unambiguity, Substance, and Consistency between the participants of varied L2 levels, and (3) the higher-level L2 students significantly outperformed their lower-level counterparts in terms of conclusion and Fair-mindedness. Important implications of these findings are discussed.

| KEYWORDS

Argumentative writing; critical thinking; second language education; second language writing

| ARTICLE INFORMATION

ACCEPTED: 02 October 2023 **PUBLISHED:** 20 October 2023 **DOI:** 10.32996/ijllt.2023.6.10.15

1. Introduction

Highly valued in higher education in the 21st century, critical thinking (CT) is one of the key academic competences tertiary students are expected to enhance for their sustainable development (Kafri, 2022; Kostoulas-Makrakis & Makrakis, 2020). For example, in China, CT is increasingly promoted in tertiary second language (L2) classes (Du & Zhang, 2022; Zhang et al., 2020).

Improvement of CT abilities and development of linguistic skills are interdependent (Li, 2016, 2023), because arguments are the primary medium for displaying CT (Shehab & Nussbaum, 2015; Nussbaum, 2021). In this sense, a growing number of university instructors of L2 academic writing (e.g., Liu & Stapleton, 2014; Lu & Xie, 2019) have implemented CT-oriented pedagogical interventions for developing their students' CT capacities demonstrated in L2 written arguments.

However, despite pedagogical efforts to improve L2 writers' CT performance in argumentative writing (e.g., Liu & Stapleton, 2014), studies (e.g., Dong & Chang, 2023) have consistently revealed CT problems in L2 argumentative essays written by university students in China. Considering that a piece of L2 argumentative writing normally comprises four parts – that is, position(s), explanations, evidence, and conclusion (Blattner & Frazier, 2002; Liu, 2014; Midgette et al., 2008) – it is necessary to explore in which part of an L2 argumentative essay students perform the worst in terms of CT. Findings about students' CT performance on each part of an argumentative essay can enhance L2 writing instructors' understanding of areas of improvement in students' argumentative writing.

Furthermore, although scholars (Dong & Chang, 2023; Floyd, 2011; Luk & Lin, 2015; Manalo & Sheppard, 2016) have pointed out negative influences of limited L2 proficiency on L2 CT performance, it is arbitrary to conclude that advanced L2 proficiency

guarantees satisfactory L2 CT performance in argumentative writing. Therefore, more research attention needs to be paid to comparing whether high-level L2 students outperform their low-level counterparts in terms of L2 CT. Research into the L2 CT performance of writers with varying L2 proficiency levels could reveal differences in the CT problems encountered by high- and low-achieving L2 learners, thus allowing practitioners of L2 education to reexamine their instructional approaches for the sustainable development of their students.

To fill the abovementioned research gaps, the present study collected L2 argumentative essays written by 65 participants from a Chinese university. The following questions were addressed in the present study:

(1) How do L2 students perform in L2 argumentative writing in terms of CT?

(2) What are the differences between higher- and lower-level L2 students in CT performance?

2. Literature Review

2.1. Theoretical Framework of CT: Paul and Elder's (2014) Elements of Reasoning and CT Intellectual Standards

Paul and Elder (2014) defined CT as "self-directed, self-disciplined, self-monitored, and self-corrective thinking" (p. 6) and held that thinking critically requires knowledge of elements of reasoning and CT intellectual standards. Table 1 displays eight elements of reasoning (Paul & Elder, 2014): purposes, questions at issue, assumptions, points of view, concepts, information, inferences, and implications. Paul and Elder (2014) emphasised that the prerequisite for being a critical thinker is to identify the eight elements of reasoning. These interconnected elements are embedded in an individual's thinking processes (Paul & Elder, 2014).

Table 1. Paul and Elder's (2014) Elements of Reasoning

Elements of reasoning	Definitions
Purposes	"Goals, objectives" (Paul & Elder, 2014, p. 118)
Questions at issue	"Problem, issue" (Paul & Elder, 2014, p. 118)
Assumptions	"Presuppositions, axioms, taking for granted" (Paul & Elder, 2014, p. 118)
Points of view	"Frames of reference, perspectives, orientations" (Paul & Elder, 2014, p. 118)
Concepts	"Theories, definitions, laws, principles, models" (Paul & Elder, 2014, p. 118)
Information	"Data, facts, reasons, observations, experiences, evidence" (Paul & Elder, 2014, p. 118)
Inferences	"Conclusions, solutions" (Paul & Elder, 2014, p. 118)
Implications	"Consequences of our reasoning" (Paul & Elder, 2014, p. 124)

Table 2 presents nine CT intellectual standards recommended by Paul and Elder (2014): clarity, precision, accuracy, breadth, logic, significance, fairness, depth, and relevance. These standards can be used for two purposes (Paul & Elder, 2014): (1) to assess individuals' thinking performance and (2) to be applied by individuals in their actual thinking processes for thinking critically.

Table 2. Paul and Elder's (2014) CT Intellectual Standards

CT intellectual standards	Definitions
Clarity	"Understandable; the meaning can be grasped" (Paul & Elder, 2014, p. 117)
Precision	"Exact to the necessary level of detail" (Paul & Elder, 2014, p. 117)
Accuracy	"Free from errors or distortions; true" (Paul & Elder, 2014, p. 117)
Breadth	"Encompassing multiple viewpoints" (Paul & Elder, 2014, p. 117)
Logic	"The parts make sense together; no contradictions" (Paul & Elder, 2014, p. 117)
Significance	"Focusing on the important; not trivial" (Paul & Elder, 2014, p. 117)
Fairness	"Justifiable; not self-serving or one-sided" (Paul & Elder, 2014, p. 117)
Depth	"Containing complexities and multiple interrelationships" (Paul & Elder, 2014, p. 117)
Relevance	"Relating to the matter at hand" (Paul & Elder, 2014, p. 117)

Widely used as theoretical frameworks in many existing studies (e.g., Dong, 2023; Lu & Xie, 2019), Paul and Elder's (2014) elements of reasoning and CT intellectual standards present a comprehensive picture of CT requirements. Therefore, the present study is theoretically grounded in Paul and Elder's (2014) elements of reasoning and CT intellectual standards.

2.2. Existing Research into Assessing CT Performance in Argumentative Writing

CT is interconnected with argumentation (Nussbaum & Kardash, 2005; Nussbaum, 2021; Shehab & Nussbaum, 2015). The assessment of CT performance in written argumentation has been addressed in various writing rubrics in existing literature (e.g., Giri & Paily, 2020; Sato, 2022). For instance, Stapleton (2001) proposed that CT can be manifested in argumentative writing through elements such as "Argument" (p.515), "Evidence" (p.517), "Recognition of opposition" (p.517), and "Fallacies" (p.518). Similarly, Nussbaum et al. (2019) assessed written CT performance by evaluating performance in reasons, evidence, counterarguments, and refutation. As seen from the operationalisations of CT in written arguments, three key features have been identified (e.g., Nussbaum et al., 2019): (1) display of a central position, (2) existence of alternative positions, and (3) high-quality evidence for justifying the positions. Accordingly, these features were considered in the development of the CT scoring guide in the present study.

In recent years, many empirical studies on assessing CT performance in argumentative writing mainly provided evidence about the roles of specific argumentation elements (e.g., counterarguments) in written CT performance (e.g., Nussbaum et al., 2019; Crossley et al., 2022). Yet little research has given a comprehensive picture of CT performance in each major part of an argumentative essay. Research into how student writers perform in different parts of an argumentative essay can reveal findings about which part of an argumentative essay needs more improvements in terms of CT.

2.3. Existing Research into Influences of Various Factors on CT Performance

Scholars (e.g., Din, 2020) have conducted numerous studies on how contextual (e.g., culture, education) or individual factors (e.g., topic familiarity, language proficiency) may influence CT performance. First, two key contextual factors, culture and education, may influence students' demonstration of CT (e.g., Atkinson, 1997; Wang & Wu, 2023; Zhao, 2020). For example, when pursuing education in Anglophone countries, East Asian students – who are often influenced by inductive patterns of reasoning, Confucianism-oriented collectivist cultural values, teacher-centred pedagogical approaches, and exam-oriented education systems (McKinley, 2013; Shi, 2006) – may have difficulty expressing their ideas according to Western norms of CT (Durkin, 2008). However, it is arbitrary to conclude that East Asian students are inferior in CT, as CT-oriented pedagogical treatments could be implemented to help them adapt to deductive patterns of reasoning (e.g., Liu & Stapleton, 2014; Zhan, 2021).

Additionally, empirical evidence has revealed influences of topic familiarity on CT performance (e.g., Indah, 2017). Indeed, lack of familiarity with topics at hand may hinder CT performance (Stapleton, 2001, 2002). These prior research findings (e.g., Stapleton, 2001) have informed the research design of the present study, in which it was decided that the topic of the writing prompt needed to be relevant to the participants' personal experience.

In terms of potential influences of language proficiency on CT performance, increasing evidence (e.g., Dong & Chang, 2023) has shown that students perform better in first-language (L1) CT tasks than in L2 ones. For example, in Floyd's (2011) study, the

participants who first carried out a specific CT task in L1 outperformed those who first carried it out in L2. More recently, Dong and Chang (2023) revealed that writers demonstrated more satisfactory CT performance in L1 argumentative writing than in L2 argumentative writing. Yet these studies did not group participants according to L2 proficiency levels, thus failing to show whether there are differences in L2 CT performance between high- and low-level L2 students.

3. Methodology

3.1 Research Purposes

The purposes of this study were to examine L2 students' CT performance in L2 argumentative writing and to explore how higher-level L2 students differ from their lower-level counterparts in terms of L2 CT performance.

3.2 Participants and Data Collection Procedures

Convenience sampling was adopted in the study to recruit the participants. The principal investigator established close relationships with several professors at a prestige Chinese university, so she was allowed to distribute information sheets to potential participants during lectures. Then, some students who agreed to participate recommended their friends to participate in this study, so the principal investigator explained the research information to these potential participants via social networking software. Finally, 65 undergraduates provided informed consent. They participated in an online demographic questionnaire survey to provide their demographic information, such as age, years of English learning, major, and College English Test Band 4 (CET-4) score.

The participants, aged between 18 and 22, had grown up and received education in mainland China. Therefore, cultural experiences (e.g., Atkinson, 1997) and educational backgrounds (e.g., Durkin, 2008) were unlikely to significantly influence the L2 CT performance of the participants. Nevertheless, their chosen fields of undergraduate study differed: 34 participants were enrolled in English language or English translation programmes, 15 in humanities or social sciences programmes, and 16 in science or engineering programmes.

Prior to data collection, all 65 participants had learnt English for a minimum of six years and had successfully passed the CET-4. The CET, an influential standardised tertiary English proficiency exam in China, assesses a test taker's proficiency in various aspects of English, including listening, reading, translation, and writing (Bai, 2020). Due to the test's high reliability and validity, CET scores serve as a significant indicator of overall English proficiency levels (Zheng & Cheng, 2008). Test takers who achieve a score of at least 550 on CET-4 are typically considered high-achieving English language learners (Xu & Liu, 2019; Zheng & Cheng, 2008). Hence, based on the CET-4 scores, there were 33 higher-level L2 students and 32 lower-level L2 students. The higher-level L2 students (M = 593.27, SD = 29.98) received significantly higher CET-4 scores than did their lower-level counterparts (M = 496.44, SD = 34.32), t (63) = -12.13, p <.01, d = 3.01, 95% CI [-112.79, -80.88].

To reconfirm the English proficiency differences between the participants of varied English levels, a 30-minute English vocabulary test named C-test (i.e., Qiu, 2017) was organised in the present study. C-test is a gap-filling test used by many researchers (e.g., Qiu, 2017) to assess their research participants' English proficiency levels. In the present study, the higher-level L2 students (Mdn = 68) significantly outperformed the lower-level ones (Mdn = 56) in the C-test, U = 172, Z = -4.68, p < .01, r = 0.58.

After determining the participants' L2 proficiency levels, the principal investigator invited all the participants to attend a 20-minute training session that introduced them to completing a computer-based English writing task. During the training session, the participants practised writing in English on their computers.

Immediately following the training session, the participants were invited to write a 300-word argumentative essay in English within 40 minutes on their computers in response to the following writing prompt: To what extent do you agree or disagree with the following statement: Students can learn more effectively through online education than through traditional classes (adapted from Nambiar, 2020, para. 1)? The participants were prohibited from accessing external resources, such as dictionaries, mobile phones, and websites. They shared computer screens for allowing the principal investigator to invigilate the writing processes.

Since the data collection occurred during a period when online education was widely implemented worldwide, the participants were already familiar with the topic at hand. Therefore, the potential influence of unfamiliarity with the topic on L2 CT performance was minimised in the present study.

After completing the writing task, the participants emailed their English writing to the principal investigator for the evaluation of their L2 CT performance.

3.3. Assessment of CT Performance

With reference to Paul and Elder (2014)'s CT intellectual standards and elements of reasoning as well as by adapting existing rubrics of CT (e.g., Wen & Liu, 2006; Dong, 2017), a scoring guide of CT (Appendix A) was used in the present study for assessing CT performance in L2 argumentative writing. The CT scoring guide consists of four criteria: Unambiguity, Fair-mindedness, Substance, and Consistency. Details of how the CT intellectual standards (Paul & Elder, 2014) informed the development of the CT criteria in the present study can be found in Table 3.

Table 3. Connections between CT criteria and CT intellectual standards (Paul & Elder, 2014)

CT criteria in the present study	CT intellectual standards (Paul & Elder, 2014)
Unambiguity: with great clarity and in detail	Clarity Precision
Fair-mindedness: with neither one-sidedness nor distortions	Breadth Fairness Accuracy
Substance: relevant, persuasive, and deep	Relevance Significance Depth
Consistency: Consistent	Logic

Unambiguity encompasses two intellectual standards: clarity and precision (Paul & Elder, 2014). Clarity is based on precise details (Paul & Elder, 2014). In the pilot stage of the present study, the principal investigator and two experienced essay raters found it difficult to differentiate clarity from precision. For example, the sentence 'online education is interesting.' is not clear enough, because more details should have been provided. The word 'interesting' may puzzle readers, because sometimes it can be a negative or neutral adjective. To make the writing unambiguous, writers need to reduce the use of vague or confusing words (Paul & Elder, 2014). Thus, clarity and precision (Paul & Elder, 2014) are incorporated into one criterion, i.e., Unambiguity.

Fair-mindedness encompasses three intellectual standards: fairness, accuracy, and breadth (Paul & Elder, 2014). According to Paul and Elder (2014), achieving fairness requires justifiable thinking, so prejudice-embedded thinking does not involve fairness. Fairness and accuracy are highly interconnected. In the pilot stage of the present study, the principal investigator along with two raters held that it is hard to assess whether contents of an argumentative essay are true or not, because the participants, who were prohibited from accessing external sources during their writing processes, could not check whether the contents in their writing involve false information (e.g., inaccurate statistics). Therefore, the present study operationalised accurate arguments as arguments that contain neither distorted nor exaggerated statements, because writers of prejudice-embedded thinking may think from a one-sided perspective and generate arguments that misrepresent the truth (Paul & Elder, 2014). For example, participants who preferred online education to traditional classes claimed that students cannot learn what they like in traditional classes. However, the reality is that universities allow students to select courses they prefer before each academic term. Thus, such distorted statements deviate from reality. In contrast, using hedging devices (e.g., may) can help the arguments meet the requirements of accuracy and fairness. Furthermore, fairness involves breadth, because the purposes of fairness and breadth are to avoid "self-serving ends" (Paul & Elder, 2014, p. 116) and to think from different perspectives (Paul & Elder, 2014). Therefore, fairness, accuracy, and breadth are combined into one criterion (i.e., Fair-mindedness).

Substance encompasses three intellectual standards: relevance, significance, and depth (Paul & Elder, 2014). To meet requirements of relevance (Paul & Elder, 2014), writers need to provide claims, explanations, evidence, and conclusions that are relevant to the topic at hand (i.e., comparing the effectiveness of online education with that of traditional classes). Paul and Elder (2014) held that relevance is the prerequisite of significance, and significance involves relevance. To satisfy requirements of significance, writers need to provide relevant responses and consider important aspects of the topic at hand without thinking superficially (Paul & Elder, 2014). Furthermore, to satisfy requirements of depth, individuals need to think "beneath the surface of an issue or problem, identify the complexities inherent in it, and then deal with those complexities in an intellectually responsible way" (Paul & Elder, 2014, p. 111). During the pilot grading processes of the research, the raters thought that depth and significance are very similar. Therefore, relevance, significance, and depth are incorporated into one criterion (i.e., Substance).

Consistency aligns with the standard of logic proposed by Paul and Elder (2014). For example, to fulfil the requirements of logic, a writer's conclusion should align with his or her thesis statement and topic sentences.

According to Paul and Elder (2014), people's thinking processes are influenced by specific purposes, questions, and assumptions, leading them to think from a particular point of view. To support a point of view, individuals often provide explanations and evidence. Lastly, a conclusion needs to be drawn (Paul & Elder, 2014).

Table 4 illustrates the relationship between the four major parts of L2 argumentative writing (e.g., Liu, 2014) and six elements of reasoning (Paul & Elder, 2014). Two elements (i.e., questions at issue; purposes) were excluded from the data analysis of the present study due to two reasons: (1) the raters in the present study found that these two elements can hardly be identified in argumentative writing and (2) existing rubrics of argumentative writing (e.g., Stapleton & Wu, 2015) seldom consider these two elements.

Table 4. Connections between Major parts of an L2 Argumentative Essay (e.g., Liu, 2014) and Elements of Reasoning (Paul & Elder, 2014)

Major parts of an L2 argumentative essay	Elements of reasoning (Paul & Elder, 2014)
Position(s): central position and alternative positions, each with main claim and supporting claims (adapted from Qin & Karabacak, 2010)	Points of view
Explanations: explanations for supporting claims (adapted from Qin & Karabacak, 2010)	Concepts Assumptions
Evidence: personal experiences, statistics, anecdotes, research findings, etc. (adapted from Qin & Karabacak, 2010)	Information
Conclusion: restatement of the position(s), inferences, and implications (Liu, 2014)	Point of view Inferences Implications

As shown in Table 4, writers' positions include a central position and alternative positions (Qin & Karabacak, 2010). Writers need to explain the positions (Paul & Elder, 2014; Qin & Karabacak, 2010). Concepts and assumptions are embedded in such explanations (Paul & Elder, 2014). The positions and explanations are justified through evidence (Paul & Elder, 2014; Qin & Karabacak, 2010). To conclude an argumentative essay, writers need to restate the positions, make inferences, and provide implications (Liu, 2014). Thus, the present study evaluated CT performance in four major parts of an L2 argumentative essay: position(s), explanations, evidence, and conclusion.

As depicted in Appendix A, each of the four CT criteria (i.e., Unambiguity, Fair-mindedness, Substance, Consistency) is assessed on a 40-point scale, resulting in a total maximum score of 160 points for overall L2 CT performance. A participant's score for each criterion was determined by evaluating his or her performance in the major parts of an L2 argumentative essay: position(s) (20%), explanations (30%), evidence (30%), and conclusion (20%).

To ensure reliability of the grading process, the principal investigator first independently coded the four major parts of each essay and graded the CT performance in each piece of writing. Three weeks after the first round of the coding and marking, the principal investigator conducted the second round of the coding and marking, during which the CT performance of each essay was assessed again. The levels of the intra-rater reliability in the grading results were calculated using the intraclass correlation coefficient (ICC). A high level of agreement is typically indicated by an ICC value above 0.75 (Koo & Li, 2016). In this study, the intra-rater reliability is very high (overall L2 CT: ICC = .99; Unambiguity: ICC = .99; Fair-mindedness: ICC = .99; Substance: ICC = .99; Consistency: ICC = 1.0).

To ensure a high level of inter-rater reliability, the principal investigator invited a doctoral researcher with six years of English-medium study experience to participate in the grading processes. The invited rater first watched an instructional video about how to code different parts of an argumentative essay. Next, he practised coding position(s), explanations, evidence, and conclusion in five argumentative essays downloaded from English learning websites. After he was familiarised with the coding, he was invited to watch another instructional video about how to evaluate the CT performance according to the CT scoring guide of the present study. Then, he assessed CT performance in randomly selected 20% of the collected essays. A very high level of inter-rater reliability was achieved in the grading of the L2 CT performance (overall L2 CT: ICC = .99; Unambiguity: ICC = .99; Fair-mindedness: ICC = .99; Substance: ICC = .99; Consistency: ICC = 1.0). Disagreements over the grading results were addressed through discussions.

3.4. Normality Tests

Shapiro–Wilk tests were conducted to evaluate whether the L2 CT data are normally distributed. The results of the Shapiro–Wilk tests are presented in Table 5. According to this table, the data are normally distributed for overall L2 CT, W (65) = .99, p = .70, and for Substance, W (65) = .96, p = .054. Hence, independent samples t-tests were performed to examine whether there were differences between higher- and lower-level L2 students in terms of overall L2 CT performance and Substance. Mann-Whitney U tests were used to assess differences in other L2 CT data (e.g., Unambiguity) between the participants of varying levels of L2 proficiency.

Table 5. Results of Shapiro–Wilk Tests

Measure	Statistic	*df*	*p*
Overall L2 CT	.99	65	.70
Position(s)	.87	65	0**
Explanations	.94	65	.003**
Evidence	.61	65	0**
Conclusion	.89	65	0**
Unambiguity	.94	65	.002**
In position(s)	.70	65	0**
In explanations	.80	65	0**
In evidence	.52	65	0**
In conclusion	.82	65	0**
Fair-mindedness	.96	65	.02*
In position(s)	.72	65	0**
In explanations	.68	65	0**
In evidence	.57	65	0**
In conclusion	.84	65	0**
Substance	.96	65	.054
In position(s)	.58	65	0**
In explanations	.68	65	0**
In evidence	.59	65	0**
In conclusion	.82	65	0**
Consistency	.84	65	0**
In position(s)	.58	65	0**
In explanations	.73	65	0**
In evidence	.52	65	0**
In conclusion	.83	65	0**

Note: Statistical significance: ** p < 0.01; * p < 0.05.

4. Results and Discussions

4.1. Results of the First Research Question

Table 6 reveals that the mean score for the overall L2 CT performance (higher L2 level: M = 45.55, SD = 9.11; lower L2 level: M = 40.44, SD = 12.09) was less than half of the full marks (i.e., full marks = 140). Consequently, the overall L2 CT performance of the participants, regardless of L2 proficiency levels, was deemed unsatisfactory.

Table 6. Mean Scores of L2 CT (By L2 Proficiency)

Measure	Higher L2 level			Lower L2 level		
	n	***M***	***SD***	***n***	***M***	***SD***
Overall L2 CT	33	45.55	9.11	32	40.44	12.09
Position(s)	33	12.91	3.88	32	11.88	3.17
Explanations	33	17.91	5.64	32	17.53	5.45
Evidence	33	3.73	6.79	32	4.03	7.17
Conclusion	33	10.73	5.22	32	7	7.33
Unambiguity	33	13.33	4.36	32	11.69	5.07
In position(s)	33	3.70	1.67	32	3.50	1.52
In explanations	33	6.09	3.48	32	5.44	2.99
In evidence	33	0.82	1.55	32	1.13	2.38
In conclusion	33	2.73	1.64	32	1.63	2.12
Fair-mindedness	33	12.33	3.45	32	10.38	4.78
In position(s)	33	4.18	2.02	32	3.31	1.31
In explanations	33	4.27	1.51	32	3.94	1.61
In evidence	33	0.91	1.76	32	0.94	1.78
In conclusion	33	2.97	1.67	32	2.19	2.56
Substance	33	8.88	2.72	32	9.03	3.06
In position(s)	33	2.24	0.83	32	2.56	0.91
In explanations	33	3.27	1.38	32	3.94	2.21
In evidence	33	1	1.79	32	1.03	1.96
In conclusion	33	2.36	1.54	32	1.50	1.61
Consistency	33	10.73	3.45	32	9.34	3.50
In position(s)	33	2.79	.99	32	2.50	1.24
In explanations	33	4.27	2.13	32	4.22	1.68
In evidence	33	1	2.33	32	0.94	1.78
In conclusion	33	2.61	1.62	32	1.69	1.97

The mean score for the overall L2 CT performance was derived from the mean scores of L2 CT in positions, explanations, evidence, and conclusion. As displayed in Table 6, the mean score in explanations (higher L2 level: M = 17.91, SD = 5.64; lower L2 level: M = 17.53, SD = 5.45) accounted for the largest proportion of the mean score for the overall L2 CT, whereas the mean score in evidence (higher L2 level: M = 3.73, SD = 6.79; lower L2 level: M = 4.03, SD = 7.17) constituted the smallest proportion of the mean score for the overall L2 CT.

Regarding the performance of each CT criterion, Table 6 reveals that the mean score for Unambiguity (higher L2 level: M = 13.33, SD = 4.36; lower L2 level: M = 11.69, SD = 5.07) was higher than the mean scores for Fair-mindedness (higher L2 level: M = 12.33, SD = 3.45; lower L2 level: M = 10.38, SD = 4.78), Substance (higher L2 level: M = 8.88, SD = 2.72; lower L2 level: M = 9.03, SD = 3.06), and Consistency (higher L2 level: M = 10.73, SD = 3.45; lower L2 level: M = 9.34, SD = 3.50). Furthermore, Table 6 shows the performance of Unambiguity in positions, explanations, evidence, and conclusion. As displayed in Table 6, the most disappointing performance of Unambiguity was reflected in evidence (higher L2 level: M = 0.82, SD = 1.55; lower L2 level: M = 1.13, SD = 2.38), whereas the most satisfactory performance of Unambiguity was evident in explanations (higher L2 level: M = 6.09, SD = 3.48; lower L2 level: M = 5.44, SD = 2.99).

Table 7 presents the frequencies of the scores of Unambiguity in explanations and evidence. Forty-eight participants scored zero in terms of Unambiguity in evidence, whereas only two participants scored zero in terms of Unambiguity in explanations. Additionally, eleven participants achieved the maximum score of 12 points for Unambiguity in explanations, whereas only one participant attained 12 points for Unambiguity in evidence.

Table 7. Frequencies of Different Scores of Unambiguity in Explanations and Evidence

Scores	Unambiguity in explanations		Unambiguity in evidence	
	Frequencies		*Frequencies*	
	Higher L2 level	Lower L2 level	Higher L2 level	Lower L2 level
0	1	1	25	23
3 points	11	12	7	8
6 points	14	15	1	0
12 points	7	4	0	1

Extract 1 provides an example of satisfactory performance in Unambiguity within explanations. In this extract, Participant A, who achieved the maximum score of 12 points for Unambiguity in explanations, elaborated on one of his or her topic sentences (i.e., *A second reason is online education provides a comfortable learning environment for students to learn at their own pace*). This participant clearly explained why students' learning can benefit more from online education than from traditional classes.

Extract 1 Satisfactory Performance of Unambiguity in Explanations (Participant A)

For different learners, their learning habits vary differently. In tradition education, students must follow a contained routine to study at certain classrooms, finish their homework in a schedule and attend the test in the end date. But with an online course, you can not only learn at you own place with a comfortable rate, but also incorporate other reference materials to supplement your learning to master the subject.

Extract 2 displays an example of unsatisfactory performance of Unambiguity in evidence. In Extract 2, Participant B briefly introduced a course from a renowned university as evidence to support one of his or her topic sentences (i.e., *By the online classes, brilliant teaching resources could be used for more students*). However, the one-sentence evidence lacked sufficient details, as it failed to explain why this course was beneficial for students' learning.

Extract 2 Unsatisfactory Performance of Unambiguity in Evidence (Participant B)

…for example, the MIT offered the Procedure Programming for all over the world students.

Table 6 shows that the mean score of Fair-mindedness in evidence (higher L2 level: M = 0.91, SD = 1.76; lower L2 level: M = 0.94, SD = 1.78) accounted for the smallest proportion of the mean score of Fair-mindedness. Conversely, the mean score of Fair-mindedness in explanations (higher L2 level: M = 4.27, SD = 1.51; lower L2 level: M = 3.94, SD = 1.61) constituted the largest proportion of the mean score of Fair-mindedness. Table 8 displays that none of the participants achieved the maximum score of 12 points for Fair-mindedness in explanations and evidence. Additionally, only one participant received a score of zero for Fair-mindedness in explanations, while 49 participants received a score of zero for Fair-mindedness in evidence.

Table 8. Frequencies of Different Scores of Fair-mindedness in Explanations and Evidence

Scores	Fair-mindedness in explanations		Fair-mindedness in evidence	
	Frequencies		*Frequencies*	
	Higher L2 level	Lower L2 level	Higher L2 level	Lower L2 level
0	0	1	25	24
3 points	19	20	6	6
6 points	14	11	2	2
12 points	0	0	0	0

Extract 3 presents an example of satisfactory performance of Fair-mindedness within a participant's explanations. This participant effectively employs a hedge (i.e., 'can') and linguistic markers of comparative forms (i.e., 'more') to mitigate degrees of certainty in the explanations. In this sense, without showing any indications of distortion, this extract serves as an example of satisfactory performance in Fair-mindedness within explanations.

Extract 3 Satisfactory Performance of Fair-mindedness in Explanations (Participant C)

…*because online education does not require face-to-face interaction, it can reduce stress and create a sense of relaxation, allowing students to feel more comfortable and flexible in class.*

Extract 4 shows an example of unsatisfactory performance in evidence regarding Fair-mindedness. The absence of hedging devices heightens levels of prejudice and reduces the degree of fairness in the evidence.

Extract 4 Unsatisfactory Performance of Fair-mindedness in Evidence (Participant D)

As far as I am concerned, my eyes become dry and exhausted, causing discomfort both physically and emotionally, and ultimately leading to lower learning efficiency.

Table 6 reveals that the mean score for Substance was predominantly derived from the mean score for Substance in explanations (higher L2 level: *M* = 3.27, *SD* = 1.38; lower L2 level: *M* = 3.94, *SD* = 2.21) but least influenced by the mean score for Substance in evidence (higher L2 level: *M* = 1, *SD* = 1.79; lower L2 level: *M* = 1.03, *SD* = 1.96). Table 9 displays that only one participant received the maximum score of 12 points for Substance in explanations. Additionally, no participants obtained the maximum score of 12 points for Substance in evidence.

Table 9. Frequencies of Different Scores of Substance in Explanations and Evidence

Scores	Substance in explanations		Substance in evidence	
	Frequencies		***Frequencies***	
	Higher L2 level	Lower L2 level	Higher L2 level	Lower L2 level
0	2	2	24	24
3 points	26	20	7	5
6 points	5	9	2	3
12 points	0	1	0	0

Extract 5 provides an example of satisfactory performance in Substance within explanations. These explanations were written to support a topic sentence regarding students' lack of learning motivation in online education. Relevant to the writing prompt about comparing learning effectiveness between online education and traditional classes, this extract provides readers with an understanding of how the study atmosphere in online education may hinder students' effective learning.

Extract 5 Satisfactory Performance of Substance in Explanations (Participant E)

Studying online typically implies that students can receive education at home without needing to go to physical classrooms. However, whether at home or in the dormitory, the atmosphere tends to be too relaxed, which is not conducive to studying. Without any motivation or encouragement from teachers and peers, students lacking discipline may not exert the same level of effort when completing homework after class. Consequently, their minds may wander, resulting in decreased attention and focus.

Extract 6 demonstrates how Participant F used personal experience as evidence to support one of his or her topic sentences (i.e., *What's more, online education includes more diverse courses than traditional classes*). However, the evidence is not sufficiently convincing, as it relies on the participant's subjective feelings and personal experience.

Extract 6 Unsatisfactory Performance of Substance in Evidence (Participant F)

Take my experience as an example: I often prefer to choose online courses to improve myself. That way, I can acquire knowledge and skills through the courses that appeal to me.

Table 6 presents that the highest level of Consistency performance was observed in explanations (higher L2 level: *M* = 4.27, *SD* = 2.13; lower L2 level: *M* = 4.22, *SD* = 1.68), while the lowest level of Consistency performance was observed in evidence (higher L2 level: *M* = 1, *SD* = 2.33; lower L2 level: *M* = 0.94, *SD* = 1.78). Table 10 presents frequencies of different scores of Consistency in explanations and evidence. Only 2 participants scored zero in terms of Consistency in explanations, while 49 participants scored zero in terms of Consistency in evidence.

Table 10. Frequencies of Different Scores of Consistency in Explanations and Evidence

Scores	Consistency in explanations		Consistency in evidence	
	Frequencies		***Frequencies***	
	Higher L2 level	Lower L2 level	Higher L2 level	Lower L2 level
0	1	1	25	24
3 points	19	17	7	6
6 points	12	14	0	2
12 points	1	0	1	0

Extract 7 demonstrates an example of satisfactory performance of Consistency within explanations. By offering support for a topic sentence (i.e., *Second, universities and schools provide better environments for students to work efficiently.*), these explanations effectively compare the differences in study environments between online education and traditional classes. As a result, the explanations align consistently with the intended meaning of the topic sentence.

Extract 7 Satisfactory Performance of Consistency in Explanations (Participant H)

In the university library, hundreds of students gather in one room to work and learn together, creating an atmosphere of collective focus and motivation. However, when studying at home, it can be challenging to maintain concentration for extended periods, especially if there are distractions like a noisy younger sibling or a nagging parent.

Extract 8 provides an example of unsatisfactory performance in Consistency within evidence. This piece of evidence was presented to support a topic sentence (i.e., *Initially, most of the online education did not require opening the video, which allowed students more flexibility for other activities, even sleeping*.). However, the evidence does not effectively convey the intended meaning of the topic sentence and fails to provide convincing justifications.

Extract 8 Unsatisfactory Performance of Consistency in Evidence (Participant I)

In my own experience, during the period of COVID-19, I would like to wake up at five to eight when I had classes at 8 o'clock.

4.2. Discussions of Key Findings in response to the First Research Question

Neither the higher-level L2 students nor the lower-level ones performed satisfactorily in terms of L2 CT. This important finding may be attributed to the prevalent focus on improving learners' basic language skills (e.g., grammar and vocabulary) in China's English as a foreign language (EFL) classes (e.g., Li, 2016; Ma & Luo, 2021). For example, Li (2016) found that more than half of her research participants, who were experienced English language instructors in China, opposed cultivating their students' CT abilities in EFL classrooms. These teachers thought that their students' development of English language skills should be the focus of their English teaching, because the improvement of their learners' CT skills may not guarantee high scores in English proficiency exams (Li, 2016). Therefore, Chinese EFL instructors are heavily influenced by high-stakes English proficiency tests (Li, 2016; Ma & Luo, 2021). They tend to focus on enhancing their students' linguistic accuracy (e.g. producing error-free sentences), fluency (e.g., producing few pauses in verbal or writing tasks) and complexity (e.g., producing sentences of syntactic complexity) (Peng et al., 2020), since large proportions of China's high-stakes English tests are comprised of sections of listening, reading, translating, speaking, and grammar (Jin & Fan, 2011; Zheng & Cheng, 2008). For instance, CET-4 consists of "listening comprehension (249 points, 35%), reading comprehension (249 points, 35%), cloze or error correction (70 points, 10%), and writing and translation (142 points, 20%)" (Zheng & Cheng, 2008, p. 409). As a result, the training of CT skills may often give way to the training of basic English language skills in tertiary EFL education in China (Zhou, 2018).

The mean score in evidence constituted the smallest proportion of the mean score for the overall L2 CT. More than 60% of the participants failed to provide any evidence in their writing. These statistical findings highlight a lack of awareness among many participants regarding the importance of justifying their arguments with evidence in argumentative writing. Additionally, the study has identified that, although 17 participants included evidence in their essays, they predominantly relied on personal experience rather than compelling scientific research findings or influential publicly available statistical data. These findings are consistent with previous research evidence (e.g., Zhang, 2018).

Three possible reasons may explain the findings regarding the lack of convincing evidence. First, this phenomenon could be attributed to the influences of Chinese traditional communication habits, which heavily rely on personal subjective judgments to persuade others (Zhang, 2018). Second, some participants may have struggled to gather sufficient compelling examples or persuasive evidence to support their arguments. Third, it is possible that existing English writing rubrics used in influential high-

stakes English language tests, which typically emphasise language proficiency, organisation, and idea development (Zhao & Huang, 2020), do not explicitly encourage test takers to give evidence in writing. Consequently, under timed test-like conditions, the participants might prioritise meeting the essay length requirement and avoiding linguistic errors to maximise their scores (Barkaoui, 2016; Porte, 1996). Allocating time to providing evidence might reduce the time available for writing a conclusion and increase the likelihood of making grammatical mistakes.

The mean score of Unambiguity was higher than the mean scores of the other three criteria of CT, with Substance receiving the lowest mean score. Achieving full marks in Unambiguity requires writers to express their positions, explanations, evidence, and conclusion with clarity and in detail. The relatively high mean score of Unambiguity may be attributed to previous L2 writing instruction that enabled many participants to understand how to structure L2 argumentative writing. Consequently, most of them included a main position, along with explanations and a conclusion, in their writing. However, attaining full marks in Substance requires writers to provide convincing and topic-relevant positions, explanations, evidence, and conclusions. It is not surprising that meeting the requirements for full marks in Substance may pose a significant challenge for L2 writers. Influenced by current large-scale English language tests that place emphasis on surface-level language abilities (e.g., vocabulary, spelling, and grammar) (Zheng & Cheng, 2008), the participants may pay less attention to content-level aspects of their English writing (e.g., relevant, convincing, and in-depth justifications for their opinions).

4.3. Results of the Second Research Question

Tables 6 and 11 reveal no significant differences between the higher-level L2 students (overall L2 CT: M = 45.55, SD = 9.11; Substance: M = 8.88, SD = 2.72) and their lower-level counterparts (overall L2 CT: M = 40.44, SD = 12.09; Substance: M = 9.03, SD = 3.06) in terms of the mean scores for overall L2 CT, t (63) =1.93 , p = .06, d = 0.48, 95% CI [-0.19, 10.40], and Substance, t (63) = -0.21, p = .83, d = 0.05, 95% CI [-1.59, 1.28].

Table 11. Results of Independent Samples T-Tests (By L2 Proficiency)

Measure	Independent samples t tests				
	t	***df***	***p***	***d***	***95% CI***
Overall L2 CT	1.93	63	.06	0.48	[-0.19, 10.40]
Substance	-0.21	63	.83	0.05	[-1.59, 1.28]

Note: Statistical significance: ** p < 0.01; * p < 0.05.

Table 12 reveals that the higher-level L2 students (Mdn = 11) significantly outperformed their lower- level counterparts (Mdn = 10) in terms of Fair-mindedness, U = 366, Z = -2.14, p = .03, r = 0.26. However, no significant different were reflected between the participants of varied L2 levels in the performance of Unambiguity, U = 385, Z = -1.89, p = .06, r = 0.23, and Consistency, U = 389, Z = -1.83, p = .07, r = 0.23.

Table 12. Results of Mann-Whitney U Tests (By L2 Proficiency)

Measure	Higher L2 level (*n* = 33)	Lower L2 level (*n* = 32)	Mann–Whitney U tests			
	Mdn	*Mdn*	*U*	*Z*	*p*	*r*
Overall	46	39.50				
Position(s)	12	12	460.50	-0.91	.36	0.11
Explanations	18	18	519.50	-0.11	.91	0.01
Evidence	0	0	519	-0.15	.88	0.02
Conclusion	10	8	374	-2.06	.04*	0.26
Unambiguity	12	10	385	-1.89	.06	0.23
In position(s)	4	4	498	-0.45	.65	0.06
In explanations	6	6	481.50	-0.66	.51	0.08
In evidence	0	0	507.50	-0.35	.73	0.04
In conclusion	2	2	309.50	-3.04	.002*	0.04
Fair-mindedness	11	10	366	-2.14	.03*	0.26
In position(s)	4	4	412.50	-1.71	.09	0.21
In explanations	3	3	476	-0.80	.42	0.21
In evidence	0	0	524	-0.07	.94	0.10
In conclusion	4	2	386.50	-1.96	.05	0.24
Substance	9	9				
In position(s)	2	2	448	-1.44	.15	0.18
In explanations	3	3	446.50	-1.34	.18	0.17
In evidence	0	0	521.50	-0.11	.91	0.01
In conclusion	2	2	374.50	-2.18	.03*	0.03
Consistency	10	8.5	389	-1.83	.07	0.23
In position(s)	2	2	425.50	-1.70	.09	0.21
In explanations	3	3	513	-0.23	.82	0.03
In evidence	0	0	521	-0.12	.90	0.02
In conclusion	2	2	360.50	-2.33	.02*	0.03

Note: Statistical significance: ** $p < 0.01$; * $p < 0.05$.

Table 12 shows that the higher-level L2 students (*Mdn* = 10) did significantly more satisfactorily than the lower-level L2 students (*Mdn* = 8) in the CT performance of conclusion, $U = 374$, $Z = -2.06$, $p = .04$, $r = 0.26$. Extract 9 displays a satisfactory conclusion written by a higher-level L2 student. The conclusion includes a restatement of the main position along with summarised supporting reasons. Additionally, the participant provides implications in the conclusion.

Extract 9 Conclusion Written by Participant J (A Higher-level L2 Student)

In conclusion, I disagree with the statement that students can learn more effectively through online education than through traditional classes. Traditional classes, which foster more interaction among students and teachers, also enhance interpersonal relationships. I firmly believe that human connection is a crucial factor for our improved performance and receiving valuable feedback in traditional classes.

Extract 10 shows a disappointing example of a conclusion written by a lower-level L2 student. In this extract, the participant merely restates the central position without summarising any supporting reasons. Furthermore, there are no significant implications in this conclusion.

Extract 10 Conclusion Written by Participant K (A Lower-level L2 Student)

In a word, students can learn more though traditional classes. Learning online is not suitable for students with poor autonomy.

4.4. Discussions of Key Findings in Response to the Second Research Question

The higher-level L2 students did not significantly outperform their lower-level counterparts in terms of overall L2 CT, Unambiguity, Substance, and Consistency. These findings indicated that high L2 proficiency levels do not guarantee satisfactory L2 CT performance. In other words, L2 learners who perform well in standardised L2 proficiency tests may still struggle with cognitively demanding L2 argumentative writing tasks. According to Cummins (1979), there are two types of linguistic skills: Basic Interpersonal Communicative Skills (BICS) and Cognitive/Academic Language Proficiency (CALP). BICS refers to surface-level linguistic skills used in informal communicative interactions, while CALP involves cognitively demanding linguistic skills, such as academic writing skills, used in formal academic settings (Cummins, 1979). Having high-level BICS does not necessarily imply having high-level CALP

(Cummins, 1979). In the present study, participants with high L2 proficiency levels may possess strong BICS, enabling them to achieve high scores on high-stakes L2 tests, such as CET-4. However, these proficient L2 learners may give little attention to the development of L2 CT skills or L2 CALP, as current standardised L2 tests tend to prioritise linguistic complexity, accuracy, and fluency in L2 outputs (Zhan & Andrews, 2014).

Significant differences were observed in performance of Fair-mindedness between the higher- and lower-level L2 students. A possible reason for this finding is that the higher-level L2 students may have focused more on improving their productive skills, particularly in writing, during their L2 learning processes. As a result, they may have developed a greater awareness of considering alternative positions in their L2 argumentative writing. In contrast, the lower-level L2 students may have prioritised such activities as vocabulary memorisation, grammar exercises, and pronunciation correction in their daily L2 learning practices, leading to a limited understanding of how to avoid one-sidedness in their argumentative writing.

The higher-level L2 students did significantly more satisfactorily than the lower-level L2 students in the CT performance of conclusion. A possible explanation for this finding is that the limited L2 proficiency may have prevented many low-achieving L2 students from writing the essay smoothly and finishing writing the conclusion within time limits. These low-achieving L2 students might have spent much time writing the introductory and body paragraphs, leaving little time for concluding their essay.

5. Conclusion

5.1. Key Findings and Implications

The first important finding is that the participants, regardless of L2 proficiency levels, failed to perform satisfactorily in terms of L2 CT. This finding indicated that L2 instructors may need to provide opportunities in classes for their students to enhance CT skills. For example, teachers can ask open-ended questions as much as possible for stimulating their learners' CT. Furthermore, instead of focusing only on language-related errors in the verbal or written outputs of their students, teachers may need to pay attention to correcting content-related problems in these outputs (Storch & Tapper, 2009).

The second important finding is that many participants failed to provide any evidence in their L2 argumentative writing. Thus, L2 instructors could consider asking their students to accumulate different types of evidence (e.g., research findings, statistics) in response to a writing topic on a weekly basis.

The third important finding is that the mean score of Substance was lower than the mean scores of the other three criteria of L2 CT. This finding implied that it is necessary for L2 writing instructors to improve their students' ability to construct persuasive written arguments in response to writing prompts. For example, when providing feedback for students' argumentative writing, instructors may need to check whether topic sentences give relevant responses to a specific writing prompt. If the topic sentences are relevant to the writing topic at hand, teachers may further need to evaluate whether persuasive explanations for the topic sentences have been provided.

The fourth important finding it that the higher-level L2 students did not significantly outperform their lower-level counterparts in terms of overall L2 CT, Unambiguity, Substance, and Consistency. This finding indicated that high scores obtained in standardised L2 tests do not ensure high-level L2 CT proficiency. Since high-achieving L2 learners have already been equipped with relatively sufficient knowledge of L2 grammar and vocabulary, there is no need for them to complete many exercises in grammar and vocabulary. Instead, public speaking, debate, and argumentative writing tasks could be arranged as much as possible for the high-achieving L2 students to train their L2 CT skills. However, since low-achieving L2 students may struggle to convey their ideas in L2, teachers could adopt translanguaging-embedded pedagogical approaches, which encourage the use of individuals' full linguistic repertoires (García & Li, 2014), to allow their low-level L2 students to express their thinking first in whatever languages they prefer. Then, teachers could help the low-achieving L2 students translate the translanguaged ideas to L2 ones.

The fifth important finding is that significantly better performance of conclusion and Fairmindedness was reflected in the higher-level L2 students' writing than in the lower-level L2 students' writing. Considering that many low-achieving L2 learners may feel it difficult to finish their writing tasks within time limits, relevant instruction about test-taking strategies (e.g., time management skills) can be provided for the less proficient students. Furthermore, the low-achieving L2 students need to be taught how to avoid one-sidedness in their argumentative writing.

5.2. Limitations and Directions for Future Research

The first limitation of the study is that it was conducted under timed test-like conditions that allowed no access to external resources. This lack of access may have hindered some participants, who may have developed an awareness of including evidence in argumentative writing, from searching for examples or evidence through reliable and accurate sources. Therefore, writing tasks that allow writers to finish writing at home within a longer period (e.g., one week) can be arranged in future similar studies.

The second limitation of the study is that it remains unknown whether the participants' unsatisfactory L2 CT performance could be partly explained by their L2 instructors' lack of understandings of L2 CT. Future research could include semi-structured interviews to explore whether L2 instructors have developed a sufficient understanding of the requirements of CT in L2 argumentative writing.

Funding: The research was funded by Postgraduate Scholarship of The University of Hong Kong.
Conflicts of Interest: The author declares no conflict of interest.
ORCID iD: 0009-0001-8995-1018

References

[1] Atkinson, D. (1997). A critical approach to critical thinking in TESOL. *TESOL Quarterly*, *31*(1), 71–94. https://doi.org/10.2307/3587975
[2] Bai, Y. (2020). The relationship of test takers' learning motivation, attitudes towards the actual test use and test performance of the College English Test in China. *Language Testing in Asia*, *10*(10), 1–18. https://doi.org/10.1186/s40468-020-00108-z
[3] Barkaoui, K. (2016). What and when second-language learners revise when responding to timed writing tasks on the computer: The roles of task type, second language proficiency, and keyboarding skills. *The Modern Language Journal*, *100*(1), 320-340. https://doi.org/10.1111/modl.12316
[4] Blattner, N. H., & Frazier, C. L. (2002). Developing a performance-based assessment of students' critical thinking skills. *Assessing Writing*, *8*(1), 47-64. https://doi.org/10.1016/S1075-2935(02)00031-4
[5] Crossley, S., Tian, Y., & Wan, Q. (2022). Argumentation features and essay quality: Exploring relationships and incidence counts. *Journal of Writing Research*, *14*(1), 1-34. https://doi.org/10.17239/jowr-2022.14.01.01
[6] Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism*, *19*, 121–129.
[7] Din, M. (2020). Evaluating university students' critical thinking ability as reflected in their critical reading skill: A study at bachelor level in Pakistan. *Thinking Skills and Creativity*, *35*, 100627. https://doi.org/10.1016/j.tsc.2020.100627
[8] Dong, Y. (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
[9] Dong, Y., & Chang, X. (2023). Investigating EFL writers' critical thinking performance across languages. *Thinking Skills and Creativity*, *47*, 1–11. https://doi.org/10.1016/j.tsc.2023.101232
[10] Du, X., & Zhang, L. (2022). Investigating EFL learners' perceptions of critical thinking learning affordances: Voices from Chinese university English majors. *Sage Open*, *12*(2), 21582440221094584.
[11] Durkin, K. (2008). The adaptation of East Asian masters students to western norms of critical thinking and argumentation in the UK. *Intercultural Education*, *19*(1), 15–27. https://doi.org/10.1080/14675980701852228
[12] Floyd, C. B. (2011). Critical thinking in a second language. *Higher Education Research and Development*, *30*(3), 289–302. https://doi.org/10.1080/07294360.2010.501076
[13] García, O., & Li, W. (2014). *Translanguaging: Language, bilingualism and education*. Palgrave. https://doi.org/10.1057/9781137385765
[14] Giri, V., & Paily, M. U. (2020). Effect of scientific argumentation on the development of critical thinking. *Science & Education*, *29*(3), 673-690. https://doi.org/10.1007/s11191-020-00120-y
[15] Indah, R. N. (2017). Critical thinking, writing performance and topic familiarity of Indonesian EFL learners. *Journal of Language Teaching and Research*, *8*(2), 229-236. http://dx.doi.org/10.17507/jltr.0802.04
[16] Jin, Y., & Fan, J. (2011). Test for English majors (TEM) in China. *Language Testing*, *28*(4), 589-596. https://doi.org/10.1177/0265532211414852
[17] Kafri, B. AL. (2022). Critical thinking (CT) in sustainable higher education: Ensuring consistent CT perception-practice and identifying gaps between college instructors' and students' perceptions in advanced academic writing courses in the UAE. *Thinking Skills and Creativity*, *46*, 1–11. https://doi.org/10.1016/j.tsc.2022.101182
[18] Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting Intraclass Correlation Coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012
[19] Kostoulas-Makrakis, N., & Makrakis, V. (2020). *Educating for a sustainable world through foreign language teaching and learning*. VGM Edusercon Limited.
[20] Li, L. (2016). Integrating thinking skills in foreign language learning: What can we learn from teachers' perspectives? *Thinking Skills and Creativity*, *22*, 273–288. https://doi.org/10.1016/j.tsc.2016.09.008
[21] Li, L. (2023). Critical thinking from the ground up: Teachers' conceptions and practice in EFL classrooms. *Teachers and Teaching: Theory and Practice*. https://doi.org/10.1080/13540602.2023.2191182
[22] Liu, F. (2014). *Investigating critical thinking in the argumentative writing of English majors at a mainland Chinese university: Implications for policy changes*. [Doctoral dissertation, The Education University of Hong Kong]. EdUHK Research Repository. https://www.lib.eduhk.hk/pure-data/pub/201715257.pdf
[23] Liu, F., & Stapleton, P. (2014). Counterargumentation and the cultivation of critical thinking in argumentative writing: Investigating washback from a high-stakes test. *System*, *45*(1), 117–128. https://doi.org/10.1016/j.system.2014.05.005
[24] Lu, D., & Xie, Y. (2019). The effects of a critical thinking oriented instructional pattern in a tertiary EFL argumentative writing course. *Higher Education Research and Development*, *38*(5), 969–984. https://doi.org/10.1080/07294360.2019.1607830
[25] 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
[26] Ma, L., & Luo, H. (2021). Chinese pre-service teachers' cognitions about cultivating critical thinking in teaching English as a foreign language. *Asia Pacific Journal of Education*, *41*(3), 543-557. https://doi.org/10.1080/02188791.2020.1793733

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

[28] McKinley, J. (2013). Displaying critical thinking in EFL academic writing: A discussion of Japanese to English contrastive rhetoric. *RELC Journal*, *44*(2), 195–208. https://doi.org/10.1177/0033688213488386

[29] Midgette, E., Haria, P., & MacArthur, C. (2008). The effects of content and audience awareness goals for revision on the persuasive essays of fifth-and eighth-grade students. *Reading and Writing*, *21*, 131-151.

[30] Nambiar, M. (2020). *Nowadays online education has become more popular as more institutes are offering courses online / Band 8 IELTS essay sample*. https://www.ielts-practice.org/nowadays-online-education-has-become-more-popular-as-more-institutes-are-offering-courses-online-band-8-ielts-essay-sample/

[31] Nussbaum, E. M., & Kardash, C. M. (2005). The effects of goal instructions and text on the generation of counterarguments during writing. *Journal of Educational Psychology*, *97*(2), 157–169. https://doi.org/10.1037/0022-0663.97.2.157

[32] Nussbaum, E. M., Dove, I. J., Slife, N., Kardash, C. A. M., Turgut, R., & Vallett, D. (2019). Using critical questions to evaluate written and oral arguments in an undergraduate general education seminar: A quasi-experimental study. *Reading and Writing*, *32*(6), 1531–1552. https://doi.org/10.1007/s11145-018-9848-3

[33] Nussbaum, E. M. (2021). Critical integrative argumentation: Toward complexity in students' thinking. *Educational Psychologist*, *56*(1), 1-17. https://doi.org/10.1080/00461520.2020.1845173

[34] Paul, R., & Elder, L. (2014). *Critical thinking: Tools for taking charge of your learning and your life* (3rd ed.). Pearson

[35] Peng, J., Wang, C., & Lu, X. (2020). Effect of the linguistic complexity of the input text on alignment, writing fluency, and writing accuracy in the continuation task. *Language Teaching Research*, *24*(3), 364–381. https://doi.org/10.1177/1362168818783341

[36] Porte, G. (1996). When writing fails: How academic context and past learning experiences shape revision. *System*, *24*(1), 107-116. https://doi.org/10.1016/0346-251X(95)00056-P

[37] Qin, J., & Karabacak, E. (2010). The analysis of Toulmin elements in Chinese EFL university argumentative writing. *System*, *38*(3), 444-456.

[38] Qiu, X. (2017). *Exploring task design, implementation procedure and second language oral performance: The effects of topic familiarity, task repetition and task types*. [Doctoral dissertation, The University of Hong Kong]. The HKU Scholars Hub. https://hub.hku.hk/handle/10722/258814

[39] Sato, T. (2022). Assessing critical thinking through L2 argumentative essays: An investigation of relevant and salient criteria from raters' perspectives. *Language Testing in Asia*, *12*(1), 1-19. https://doi.org/10.1186/s40468-022-00159-4

[40] Shehab, H. M., & Nussbaum, E. M. (2015). Cognitive load of critical thinking strategies. *Learning and Instruction*, *35*, 51–61. https://doi.org/10.1016/j.learninstruc.2014.09.004

[41] Shi, L. (2006). The successors to Confucianism or a new generation? A questionnaire study on Chinese students' culture of learning English. *Language, Culture and Curriculum*, *19*(1), 122–147. https://doi.org/10.1080/07908310608668758

[42] Stapleton, P. (2001). Assessing critical thinking in the writing of Japanese university students. *Written Communication*, *18*(4), 506–548.

[43] Stapleton, P. (2002). Critical thinking in Japanese L2 writing: Rethinking tired constructs. *ELT Journal*, *56*(3), 250–257. https://doi.org/10.1093/elt/56.3.250

[44] Storch, N., & Tapper, J. (2009). The impact of an EAP course on postgraduate writing. *Journal of English for Academic Purposes*, *8*(3), 207-223. https://doi.org/10.1016/j.jeap.2009.03.001

[45] Wang, Y., & Wu, Z. (2023). Adapting or adopting? Critical thinking education in the East Asian cultural sphere: A systematic integrative review. *Thinking Skills and Creativity*, 101330. https://doi.org/10.1016/j.tsc.2023.101330

[46] Wen, Q., & Liu, R. (2006). cóng yīng yǔ yì lùn wén fēn xī dà xué shēng chōu xiàng sī wéi tè diǎn [An exploratory study on features on English majors' abstract thinking in English argumentative compositions]. *Journal of Foreign Languages*, 162. 49-58

[47] Xu, J., & Liu, Y. (2019). CET-4 score analysis based on data mining technology. *Cluster Computing*, *22*, 3583–3593. https://doi.org/10.1007/s10586-018-2208-x

[48] Zhan, Y., & Andrews, S. (2014). Washback effects from a high-stakes examination on out-of-class English learning: Insights from possible self theories. *Assessment in Education: Principles, Policy and Practice*, *21*(1), 71–89. https://doi.org/10.1080/0969594X.2012.757546

[49] Zhan, Y. (2021). What matters in design? Cultivating undergraduates' critical thinking through online peer assessment in a Confucian heritage context. *Assessment & Evaluation in higher education*, *46*(4), 615-630. https://doi.org/10.1080/02602938.2020.1804826

[50] Zhang, Y. (2018). An investigation into the development of structure and evidence use in argumentative writing. *Theory and Practice in Language Studies*, *8*(11), 1441. https://doi.org/10.17507/tpls.0811.08

[51] Zhang, H., Yuan, R., & He, X. (2020). Investigating university EFL teachers' perceptions of critical thinking and its teaching: Voices from China. *The Asia-Pacific Education Researcher*, *29*, 483-493.

[52] Zhao, W. (2020). Epistemological flashpoint in China's classroom reform: (How) can a 'Confucian do-after-me pedagogy' cultivate critical thinking?. *Journal of Curriculum Studies*, *52*(1), 101-117. https://doi.org/10.1080/00220272.2019.1641844

[53] Zhao, C., & Huang, J. (2020). The impact of the scoring system of a large-scale standardized EFL writing assessment on its score variability and reliability: Implications for assessment policy makers. *Studies in Educational Evaluation*, *67*, 100911. https://doi.org/10.1016/j.stueduc.2020.100911

[54] Zheng, Y., & Cheng, L. (2008). Test review: College English Test (CET) in China. *Language Testing*, *25*(3), 408–417. https://doi.org/10.1177/0265532208092433

[55] Zhou, Z. (2018). A study on the cultivation of critical thinking ability of English majors. *Theory and Practice in Language Studies*, *8*(3), 349. https://doi.org/10.17507/tpls.0803.11

Appendix A Critical Thinking Scoring Guide

Levels	*Unambiguity*	*Fair-Mindedness*	*Substance*	*Consistency*
Level 1: 40 points (Central position: 20%; Explanations: 30%; Evidence: 30%; Conclusion: 20%)	**With great clarity and in detail** **Central position (8 points):** Presents the central position by articulating its main claim and supporting claim(s) in a very clear and detailed manner. **Explanations (12 points):** Explains all supporting claim(s) of the central position with great clarity and in detail. **Evidence (12 points):** Provides adequate evidence for all supporting claim(s) of the central position in a clear and detailed manner. **Conclusion (8 points):** Summarises the main claim and supporting claim(s) clearly in the conclusion and provides meaningful implications.	**With neither one-sidedness nor distortions** **Position(s) (8 points):** Presents the position(s) with neither one-sidedness nor distortions. **Explanations (12 points):** Explains the position(s) with neither one-sidedness nor distortions. **Evidence (12 points):** Provides the evidence with neither one-sidedness nor distortions. **Conclusion (8 points):** Presents the conclusion with neither one-sidedness nor distortions.	**Relevant, persuasive, and deep** **Position(s) (8 points):** Presents claim(s) that are all relevant to the task and highlight important aspects of the topic at hand. **Explanations (12 points):** Provides relevant and thorough explanations for all the supporting claim(s). **Evidence (12 points):** Gives evidence that is relevant to the task and strongly justifies the claim(s). **Conclusion (8 points):** Makes relevant inferences and provides deep implications in the conclusion.	**Consistent** **Position(s) (8 points):** Presents the supporting claim(s) that all fit together with the main claim. **Explanations (12 points):** Explains all the supporting claim(s) in a manner consistent with their intended meanings. **Evidence (12 points):** Provides evidence that can completely justify the claim(s). **Conclusion (8 points):** Draws a conclusion that is fully consistent with the central position and supported by the justifications provided.
Level 2: 20 points (Central position: 20%; Explanations: 30%; Evidence: 30%; Conclusion: 20%)	**With a few inadequacies** **Central position (4 points):** Presents the central position through either its main claim or its supporting claim(s) in a clear and detailed manner. **Explanations (6 points):** Explains the supporting claim(s) of the central position with a few inadequacies. **Evidence (6 points):** Provides some	**With either one-sidedness or distortions** **Position(s) (4 points):** Presents the position(s) with either one-sidedness or distortions. **Explanations (6 points):** Explains the position(s) with either one-sidedness or distortions. **Evidence (6 points):** Provides the evidence with either one-sidedness or distortions. **Conclusion (4 points):** Shows either one-sidedness	**Relevant but superficial** **Position(s) (4 points):** Presents the claim(s) that are all or mostly relevant to the task but may reveal superficial aspects of the main issue at hand. **Explanations (6 points):** Provides relevant but superficial explanation(s) for all or most of the supporting claim(s). **Evidence (6 points):** Gives evidence that is all or mostly relevant to the task but not convincing enough.	**Somewhat inconsistent** **Position(s) (4 points):** Presents the supporting claim(s) that mostly fit with the main claim. **Explanations (6 points):** Explains the supporting claims in ways that are mostly consistent with their intended meanings. **Evidence (6 points):** Provides evidence that can mostly justify the position(s).

	evidence for the central position. **Conclusion (4 points):** Summarises the main claim and supporting claim(s) in the conclusion but provides no implications in the conclusion.	or distortions in the conclusion.	**Conclusion (4 points):** Makes relevant inferences but provides superficial or no implications in the conclusion.	**Conclusion (4 points):** Draws a conclusion that is mostly consistent with the central position and is largely based on the justifications for the central position.
Level 3: 10 points (Central position: 20%; Explanations: 30%; Evidence: 30%; Conclusion: 20%)	**With lots of inadequacies** **Central position (2 points):** Provides few details about the central position. **Explanations (3 points):** Explains most of the supporting claim(s) of the central position ambiguously and insufficiently. **Evidence (3 points):** Provides little evidence for the central position. **Conclusion (2 points):** Summarises the main claim in the conclusion but fails to provide implications / Provides implications in the conclusion but summarises neither the main claim nor the supporting claim(s).	**With both one-sidedness and distortions** **Position(s) (2 points):** Presents the position(s) from one single perspective with distortions. **Explanations (3 points):** Explains the position(s) from one single perspective with distortions. **Evidence (3 points):** Provides the evidence from one single perspective with distortions. **Conclusion (2 points):** Shows both one-sidedness and distortions in the conclusion.	**Mostly irrelevant** **Position(s) (2 points):** Presents few relevant position(s). **Explanations (3 points):** Provides few relevant explanations for all or most of the supporting claim(s). **Evidence (3 points):** Gives little relevant evidence. **Conclusion (2 points):** Makes few relevant inferences in the conclusion.	**Mostly inconsistent** **Position(s) (2 points):** Presents supporting claims that are mostly unable to fit with their main claim. **Explanations (3 points):** Explains the supporting claims in ways that are mostly inconsistent with their intended meanings. **Evidence (3 points):** Provides little evidence that can justify the positions. **Conclusion (2 points):** Draws a conclusion that is mostly inconsistent with the central position and based on previous unjustified beliefs.
Level 4: 0 point	**Without positions, explanations, evidence, and conclusion** **Central position:** Presents no position. **Explanations:** Provides no explanation for the supporting claims of the central position.	**Without positions, explanations, evidence, and conclusion** **Position(s):** Presents no position. **Explanations:** Explains no position. **Evidence:** Provides no evidence.	**Totally irrelevant** **Position(s):** Presents no relevant position. **Explanations:** Provides no relevant explanation for the position. **Evidence:** Gives no relevant evidence. **Conclusion:** Draws no relevant conclusion.	**Totally inconsistent** **Position(s):** Presents supporting claim(s) that cannot fit with the main claim. **Explanations:** Explains the supporting claim(s) in ways that are totally inconsistent with their intended meanings.

	Evidence: Provides no evidence for the central position. **Conclusion:** Draws no conclusion.	**Conclusion:** Draws no conclusion.	**Evidence:** Provides no evidence that can justify the positions. **Conclusion:** Draws a conclusion that is totally inconsistent with the central position and based on previous unjustified beliefs.