Towards a Valid and Reliable Checklist to Evaluate Argumentative Essays Composed by ChatGPT

Danijela Ljubojević[1], Djordje M. Kadijevich[1] and Nikoleta Gutvajn[1]

[1] *Institute for Educational Research, 11/III Dobrinjska, 11000 Belgrade, Republic of Serbia*

Abstract

Since its launch a year ago, ChatGPT has sparked many concerns in education, especially when it comes to writing. Many students enjoy the benefits of getting generated text for their homework assignments; however, this behaviour impacts profoundly the writing process and the development of critical thinking skills. Among these assignments that are particularly important to critical skills development are so-called argumentative essays, which require the student to investigate a topic, collect, generate, and evaluate evidence, and establish a position on the topic in a concise manner. To assess these essays in a thoughtful way, this paper presents a checklist whose indicators focus on main aspects of essay organisation and higher-order critical thinking skills. The checklist was developed for both machine and human responses by using relevant theoretical framework (the Classical model of Argumentation and Paul-Elder critical thinking framework), the five-paragraph approach, and Cambridge English Qualifications scales at level C1 of the CEFR. As this assessment tool was applied in evaluating ChatGPT-composed argumentative essays, apart from the validity of the tool, this paper also presents its inter-rater reliability. Suggestions for research and practice are included.

Keywords

ELT, argumentative essay, writing, ChatGPT, checklist, critical thinking, reliability, validity

1. Introduction

Artificial intelligence (AI) has become an integral part of learning and teaching in many fields and English language teaching (ELT) is not an exception. Recent studies have recognised the importance of using chatbots, such as ChatGPT, in ELT [1] [2]. ChatGPT generates human-like text based on the input it receives and up to now users have been very satisfied with it. However, essay writing in English is about not only generating written content but also demonstrating good essay organization and paragraph structuring.

This pilot study set out to analyse argumentative essays generated through ChatGPT and, consequently, to come up with a reliable and valid instrument to assess students writing tasks in English. Furthermore, it aims to determine what can ChatGPT generate in terms of developing critical thinking skills. To assess the extent to which writing proficiency has been achieved, an appropriate checklist with good validity and reliability needs to be applied. Hence, this study examined validity and reliability of the developed checklist.

The findings of this research hold implications for teaching writing and the integration of AI in English language classrooms. A positive answer to the question of validity and reliability could contribute to the improvement concerning the missing instrument that assesses the promotion of critical thinking skills and the use of AI-generated content.

Proceedings for the 14th International Conference on eLearning (eLearning-2023), September 28–29, 2023, Belgrade, Serbia
EMAIL: danijela.ljubojevic@ipi.ac.rs (A. 1); djkadijevic@ipi.ac.rs (A. 2); gutvajnnikoleta@gmail.com (A. 3)
ORCID: 0000-0002-4168-7107 (A. 1); 0000-0002-5628-3455 (A. 2); 0000-0002-4337-4884 (A. 3)
© 2023 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3696/article_9.pdf

2. Theoretical Framework

Argumentative essays are types of essays which require the student to investigate a topic, collect, generate, and evaluate evidence, and establish a position on the topic in a concise manner. They are very important for developing critical thinking skills. Providing sufficient and sound arguments in the argumentative essays is essential to their success. It is not enough just to support the idea with enough details and examples; there are some more aspect that should be covered.

Developing good argumentation is vital to argumentative essays. There are different models of argumentation that be used: the Classical Model of Argumentation, the Toulmin Model, Rogerian Argumentation Model, etc. One of the most applied models in ELT is the Classical. It is also called Aristotelian because it was first mentioned in Aristotle's work *Rhetoric*. Aristotle's central idea is that persuasion comes about through arguments, i.e. by proving that something is the case. The classical argument is made up of five components, which are commonly composed in the following order [3]: introduction, narration, confirmation, refutation, and conclusion. When using this model, the writer should start with a clear, concise, and defined thesis statement that occurs in the first paragraph of the essay. Each paragraph should develop only one idea (paragraph unity) which must be supported by sufficient supporting details. What is important for this model is the use of "opposing" point of view: argumentative essays should also consider and explain differing points of view regarding the topic and discuss conflicting opinions on the topic. It is also important to use clear and logical transitions between the introduction, body, and conclusion, because without logical progression of thought, the reader will be unable to follow the essay's argument, and the structure will collapse.

A common outline for writing an argumentative essay is the five-paragraph approach (also known as the "hamburger essay," the "one-three-one essay," and the "three-tier essay."). It consists of an introductory paragraph, three body paragraphs with evidence that include discussion of opposing views, and a conclusion.

Writing argumentative essays is crucial for the development of critical thinking (CT) skills with students. CT refers to the ability to analyse and evaluate arguments or evidence. The National Council for Excellence in Critical Thinking defined it as 'the intellectually disciplined process of actively and skillfully conceptualizing, applying, analyzing, synthesizing, and/or evaluating information gathered from, or generated by, observation, experience, reflection, reasoning, or communication, as a guide to belief and action" [4]. However, it is a skill that cannot be measured directly; instead, intellectual standards are used to determine the quality of reasoning [5]. One of the critical thinking models that can be adopted in improving argumentation skills is the Paul-Elder critical thinking framework [6] [7]. The intellectual standards proposed by the framework are clarity, precision, accuracy, depth, breadth, logic, significance, relevance, and fairness and they are used for the checklist in this study.

When designing a checklist, the items addressed all five Aristotelian components and Paul-Edler's intellectual standards.

3. Methodology

The aim of this study was to determine if a reliable and valid checklist can be applied in assessing ChatGPT generated essays. It set out to explore inter-rater agreement between the two reviewers who applied the checklist developed. Furthermore, it explored if ChatGPT offers some possibilities for language teaching and learning.

In order to do this a comprehensive checklist for assessing essay structure and intellectual standards was designed. Two independent secondary school teachers graded the essays using the proposed checklist. The reviewers were not aware of the fact that they were assessing computer-generated essays.

The following instructions were given as prompts to ChatGPT:

Higher education increases the chances of employment. Agree or disagree with this statement. Support your opinion with reasons and examples. Write an essay in around 240 - 280 words.

The first researcher in this study generated the essays. The essays were then sent to the reviewers for assessment.

Essay organization was assessed using 14-item instrument, whose indicators were derived from the five-paragraph approach, the Classical model, and Cambridge English Qualifications scales at level C1 of the CEFR. The checklist was meticulously designed based on the previous research by the first author [8]. Each item gets different marks. These indicators are listed in Table 1. Ranging from 0 (lowest) to 5 (highest) and precise instructions were given for the band in the right column.

Table 1

Checklist for Essay Structure

Indicator	Items for assessment	Max. no. of points
1	Does the essay have an introduction, a body, and a conclusion?	3
2	Is the response of appropriate length?	1
	Introduction	
3	Do the general statements give background information?	1
4	Is it a funnel introduction?	1
5	Does the thesis statement state a clearly focused main idea for the whole essay?	1
	Body	
6	Are there arguments expressing the writer's point of view?	2
7	Are there arguments expressing the opposing point of view?	2
8	Does each body paragraph have a clearly stated topic sentence with a main (controlling) idea?	3
9	Does each body paragraph have good development with sufficient supporting details (facts, examples, and quotations)?	3
10	Does each body paragraph have unity (one idea per paragraph, there are no sentences that are "off the topic")?	3
11	Does each body paragraph have coherence (logical organization, transition words, and consistent pronouns)?	3
	Conclusion	
12	Does the conclusion restate your thesis or summarize your main points? What kind of conclusion does the essay have? Is it summary of the main points or restatement of the thesis?	1
13	Does the conclusion give writer’s personal opinion about the topic?	1
14	**Language (choose only one)**	5
	Uses a (wide) range of vocabulary, including less common lexis, effectively and precisely. Uses a wide range of simple and complex grammatical forms with full control, flexibility and sophistication. Errors, if present, are related to less common words and structures, or occur as slips.	*5*
	Performance shares features of Bands 3 and 5.	*4*
	Uses a range of vocabulary, including less common lexis, appropriately. Uses a range of simple and complex grammatical forms with control and flexibility. Occasional errors may be present but do not impede communication.	*3*
	Performance shares features of Bands 1 and 3.	*2*
	Uses a range of everyday vocabulary appropriately, with occasional inappropriate use of less common lexis. Uses a range of simple and some complex grammatical forms with a good	*1*

degree of control.
Errors do not impede communication.

The attainment of critical thinking skills was examined using a 9-item instrument (Table 2), whose indicators were derived from the above-mentioned Paul-Elder CT Model [9] with clarification as given by Inoshita et al. [10]. These indicators are listed in Table 2. Each item gets marks from 0 (lowest) to 3 (highest).

Table 2

Checklist for Intellectual Standards

Intellectual Standards (Paul – Elder Model) [7]	Clarification [10]
Indicator 15 Clarity	
Could you elaborate? Could you illustrate what you mean? Could you give me an example?	An essay is clear, it’s understandable and communicates information to readers with ease. None of the statements are confusing or ambiguous. There aren’t areas within the essay where the meaning is lost due to exaggerated narrative or forced and unnatural word choice. When an essay is clear, readers can follow the path that the writer is communicating. They can read smoothly without stopping to ponder what a word or even an entire sentence means.
Indicator 16 Accuracy	
How could we check on that? How could we find out if that is true? How could we verify or test that?	Is it correct? Is it true? Accuracy, not only when it comes to spelling, punctuation, and word usage, but also grammar, syntax, and conducting research within and outside of the respective disciplines.
Indicator 17 Precision	
Could you be more specific? Could you give me more details? Could you be more exact?	Precision within writing demands that words are not only spelled correctly but that their meanings are also clear and that the words are not overused. Punctuation needs to be used in a manner that follows standard rules, and ideas must be expressed in ways that are direct while still allowing for the writer to perform with skill and artistry.
Indicator 18 Relevance	
How does that relate to the problem? How does that bear on the question? How does that help us with the issue?	Is it essential to the main idea? If paragraphs in an essay are relevant, they are related to the main topic and help support the main idea with additional, related, relevant details and evidence. If paragraphs are irrelevant, a reader might think, “Wait. What? How is this on topic?” Does this point help readers understand the main issue? Does this essay focus on the assignment question or prompt? Does it answer the main question? If this paragraph is slightly off-topic, what can be done to refocus it so that it does its job in supporting the main idea in the thesis statement? If a point is confusing readers who don’t understand how it’s related to the main idea, does it belong in this essay?
Indicator 19 Depth	
What factors make this difficult? What are some of the complexities of this question? What are some of the difficulties we need to deal with?	Is it sufficiently complex? How deeply does this essay go into its topic? Is it detailed enough? Did it go far enough into the research and reviews of other texts to demonstrate a deep knowledge about the subject? How thoroughly have specific subtopics within a major been researched?

Indicator 20 Breadth	
Do we need to look at this from another perspective? Do we need to consider another point of view? Do we need to look at this in other ways?	Are all views considered? - a writer must consider not only one point of view, but all the multiple major perspectives about an issue. Is the content of an essay sufficiently comprehensive enough to cover a wide range of perspectives and angles on a given topic? Is anything missing that should be included in the scope of the topic and which would help the essay achieve enough breadth? Has the opposing view (i.e., the "naysayer's" perspective) been explored so as to strengthen the writer's own argument? (This consideration is particularly key in the development of a fully supported and wisely composed persuasive or argumentative essay.) What has not yet been considered to make this idea or essay complete?
Indicator 21 Logic	
Does all of this make sense together? Does your first paragraph fit in with your last one? Does what you say follow from the evidence?	Does a sentence, paragraph, or argument make sense? Does one point follow another point with reason and connected ideas and transitions, rather than jumping from point A to point Z without sufficient explanation of how they are related? Are all assertions fortified by sufficient evidence? Does all data collected, whether determined as relevant or not, follow a logical approach?
Indicator 22 Significance	
Is this the most important problem to consider? Is this the central idea to focus on? Which of these facts are most important?	Is the information important enough to include? Does the information answer the question asked in an assignment? What key points are most important for writers to include and for readers to consider? Is this topic important in relation to other topics in the same subject area? What's the most important thing to focus on? Is this topic worth the writer's time researching and the reader's time reading? Would there be enough readers interested in this topic?
Indicator 23 Fairness	
Is my thinking justifiable in context? Am I taking into account the thinking of others? Is my purpose fair given the situation? Am I using my concepts in keeping with educated usage, or am I distorting them to get what I want?	Does the writer of this essay exhibit the ability to fairly assess the viewpoints of others, even opposing viewpoints? Are there any fallacies, such as *ad hominems* that unfairly label opponents rather than speak directly and precisely about the opposing argument or point of view itself? Does the writer or speaker have a conflict of interest? Does that conflict of interest appear as bias in the text?

For each essay (there were 23 instances), the average Spearman correlations among the reviewers' score was calculated. These values were 0.885, 0.733, 0.754, and 0.918, respectively. As being above 0.60, these figures evidence a good agreement among the four raters, especially for the first and the last essay. Note that in many cases different measures of inter-rater reliability result in the same conclusion [11], meaning that a standard reliability threshold (>0.60) used for the kappa measure [12] could be applied.

4. Results

When the prompts were given to ChatGPT, it generated the following two essays:

Essay 1

In today's rapidly evolving job market, the role of higher education in increasing employment prospects is a topic of considerable debate. While some argue that a college degree significantly enhances one's chances of securing a job, others contend that it is not always a guarantee of employment. In this essay, I will argue that higher education does indeed increase the chances of employment, but its effectiveness depends on various factors.

To begin with, higher education equips individuals with valuable skills and knowledge that are often required in the modern workforce. Many professions, such as engineering, medicine, and law, demand specialized training and qualifications that can only be acquired through higher education. Without these credentials, individuals may find it challenging to access certain career paths or secure well-paying jobs.

Moreover, a college degree demonstrates to employers a candidate's commitment, discipline, and ability to complete a long-term educational program. It serves as a signal of a person's willingness to invest time and effort in self-improvement, qualities highly valued in the job market.

However, it's essential to acknowledge that the relationship between higher education and employment is not always straightforward. Factors such as the choice of major, the prestige of the institution, and the individual's networking skills also play significant roles. For example, a degree in computer science from a top-tier university may open doors to lucrative job opportunities, while a less specialized degree from a less-known institution might not yield the same results.

In conclusion, higher education undeniably enhances the chances of employment by providing essential skills, demonstrating commitment, and opening doors to specialized professions. However, the extent of its impact varies depending on the field of study, the institution, and the individual's networking abilities. Therefore, while higher education is a valuable asset in the pursuit of employment, it should be considered as part of a broader strategy that includes factors beyond academic qualifications.

Word count: 312

Essay 2

I strongly agree that higher education increases the chances of employment. This belief is grounded in the fact that higher education equips individuals with valuable skills, knowledge, and qualifications that are highly sought after by employers.

Firstly, higher education provides students with specialized skills and expertise that are often essential for many professions. For example, a degree in engineering equips graduates with the technical knowledge required for jobs in the engineering field. Similarly, a medical degree is necessary to become a doctor. Employers often prioritize candidates with such specialized knowledge and skills, making higher education a key factor in securing employment.

Secondly, higher education institutions offer students opportunities for personal and professional growth. Students learn critical thinking, problem-solving, communication, and teamwork skills that are transferable across various industries. These skills not only make individuals more employable but also help them succeed and advance in their chosen careers.

Furthermore, statistics consistently show a strong correlation between higher education attainment and employment rates. According to data from the U.S. Bureau of Labor Statistics, individuals with higher levels of education tend to have lower unemployment rates and higher earning potential compared to those with only a high school diploma.

For instance, in the United States, the unemployment rate for individuals with a bachelor's degree is significantly lower than for those with only a high school diploma. This demonstrates that higher education not only increases the chances of finding employment but also provides better job security.

In conclusion, higher education is a significant factor in increasing the chances of employment. It equips individuals with specialized skills, fosters personal and professional growth, and is strongly correlated with lower unemployment rates and higher earning potential. While there are exceptions, overall, pursuing higher education is a wise investment in one's future career prospects.

Word count: 296 words

These essays were marked in the following way by the reviewers (Table 3):

Table 3

Grading results for Essay 1 and Essay 2 by Reviewer 1 and Reviewer 2

Indicator	E1 R1	E1 R2	E2 R1	E2 R2
1	**3**	**3**	**3**	**3**
2	**1**	**1**	**1**	**1**
3	**1**	**1**	**1**	**1**
4	**1**	**1**	**0**	**1**
5	**1**	**1**	**1**	**1**
6	**2**	**2**	**2**	**2**
7	**2**	**1**	**2**	**0**
8	**3**	**3**	**2**	**3**
9	**3**	**3**	**3**	**3**
10	**3**	**3**	**3**	**3**
11	**3**	**3**	**3**	**3**
12	**1**	**1**	**1**	**1**
13	**1**	**0**	**1**	**0**
14	**5**	**4**	**4**	**4**
15	**3**	**3**	**3**	**3**
16	**3**	**3**	**3**	**2**
17	**3**	**3**	**2**	**3**
18	**3**	**3**	**3**	**3**
19	**3**	**2**	**3**	**2**
20	**3**	**2**	**2**	**2**
21	**3**	**3**	**3**	**3**
22	**3**	**3**	**3**	**3**
23	**3**	**2**	**2**	**3**

Reliability of this checklist was examined using inter-rater reliability based upon Spearman's correlation. This correlation suitable for ordinal data was determined using an online calculator available at https://www.socscistatistics.com/tests/spearman/default2.aspx). For the first essay, this correlation was 0.90. For the second essay, the correlation was 0.78. These figures evidence a good agreement between the two raters, especially for the first essay. Note that in many cases different measures of inter-rater reliability result in the same conclusion [11], meaning that a standard reliability threshold (>0.60) used for the kappa measure [12] could be applied.

5. Discussion

The present study was designed to determine if a reliable and valid checklist can be applied in assessing ChatGPT generated essays, to explore inter-rater agreement between the two reviewers who applied the checklist developed and explore the potential of using ChatGPT generated essays in the classroom.

It was shown that ChatGPT can produce argumentative essays that are given high marks in almost every aspect regarding the requirements set in the checklist. The results for reliability of the given instrument clearly show that the applied checklist had good psychometric features, which answers the applied research question in a positive way. It can be thus said that this checklist successfully measures one underlying construct and thus it can confidently be used in further research. Hence, the outcome of this study contributes to developing an instrument that assesses the promotion of critical thinking with the use of ChatGPT, which has been a neglected research area so far, to the authors' readings.

The results for reliability of the given instrument were affirmative, meaning that it enables a consistent, reliable assessment. Hence, the outcome of this study contributes to developing a valid and reliable instrument that assesses the promotion of critical thinking in a broader context (both for a human and machine generated responses), which has been a neglected research area so far, to the authors' readings.

6. Closing Remarks

This study set out to determine if a reliable and valid instrument can be used to assess argumentative essays in terms of essay organization and the elements of critical thinking. These essays were ChatGPT generated. The checklist was designed for this purpose comprising the requirements stemming from: the five-paragraph approach, the Classical model of argumentation, and Paul-Elder critical thinking framework. The findings have shown that the checklist is both reliable and valid so it can be used within ELT classrooms. Implications for ELT are numerous: teachers can use this checklist not only to evaluate students' argumentative essays but also to benefit from it by analysing together with students ChatGPT generated essays and thus, focus on the promotion of critical thinking skills.

Limitations of this study can be regarded in terms of the sample of essays used and the number of teachers who participated as reviewers. This limitation means that study findings need to be interpreted cautiously.

Up to now, no studies were undertaken longitudinally because ChatGPT is a new technology. Further research should be undertaken to find good learning models with the help of ChatGPT and how to implement them within educational settings.

7. Acknowledgements

The authors of this paper would like to thank the following teachers for their participation in the research as reviewers: Nana Shavishvili, Professional English Instructor, Georgian American University, Georgia; and Biljana Pipović, English Language Teacher, Grammar School "Stevan Jakovljević" Vlasotince, Serbia.

This research was funded by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Contract No. 451-03-47/2023-01/ 200018).

8. References

[1] W. C. H. Hong, “The impact of ChatGPT on foreign language teaching and learning: Opportunities in education and research,” *Journal of Educational Technology and Innovation(JETI),* vol. 5, no. 1, 2023.

[2] W. Huang, K. F. Hew and L. K. Fryer, “Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning,” *Journal of Computer Assisted Learning,* vol. 38, no. 1, p. 237–257, 2022.

[3] Aristotel, Retorika, Beograd: Štampar Makarije, 2017.

[4] National Council for Excellence in Critical Thinking, “The Foundation for Critical Thinking,” 1987. [Online]. Available: https://www.criticalthinking.org/pages/defining-critical-thinking/766.

[5] Z. S. Nakrowi, D. S. Ansori, Y. Mulyati and Y. Setyaningsih, “The use of intellectual standards to assess the quality of students’ argumentative writings,” *LITERA,* vol. 22, no. 2, pp. 200-212, 2023.

[6] L. Elder and R. Paul, The Thinker’s Guide to Analytic Thinking: How to Take Thinking Apart and What to Look for When You Do, 2nd Edition, Tomales: Rowman & Littlefield Publishers / The Foundation for Critical Thinking, 2016, p. 9.

[7] E. Linda and R. Paul, “Critical Thinking: Intellectual Standards essential to Reasoning Well Within Every Domain of Thought,” *Journal of Developmental Education,* vol. 36, no. 3, pp. 34-35, 2013.

[8] D. Ljubojevic, *Developing academic writing skills in English as L2 by means of collaborative e-learning tools,* University of Belgrade, Faculty of Philology, 2017.

[9] R. Paul and L. Elder, “Critical Thinking: Intellectual Standards Essential to Reasoning Well Within Every Domain of Human Thought, Part Two,” *Journal of Developmental Education,* vol. 37, no. 1, p. 32–36, 2013.

[10] A. Inoshita, Garland, K. K. Sims, T. Keuma, J. K. and T. Williams, English Composition - Connect, Collaborate, Communicate, Honolulu: University of Hawai, OER , 2019.

[11] A. de Raadt, M. Warrens, R. Bosker and H. A. L. Kiers, “A Comparison of Reliability Coefficients for Ordinal Rating Scales,” *Journal of Classification,* vol. 38, p. 519–543 , 2021.

[12] J. Flo, B. Landmark, O. E. Hatlevik and L. Fagerström, “Using a new interrater reliability method to test the modified Oulu Patient Classification instrument in home health care,” *Nursing Open.,* vol. 5, p. 167–175, 2018.