

Evaluating ChatGPT-4 as an automated essay scoring tool: A comparative study with human raters in Greek upper secondary education

Koskinas, Emmanouil

Democritus University of Thrace, Greece (koskinas.manolis@gmail.com)

Dosi, Ifigeneia

Democritus University of Thrace, Greece (idosi@hs.duth.gr)

Chadjipapa, Elina ✉

Democritus University of Thrace, Greece (elina@hotmail.com)

Gavriilidou, Zoe

Democritus University of Thrace, Greece (zoegab@otenet.gr)



ISSN: 2243-7703
Online ISSN: 2243-7711

OPEN ACCESS

Received: 28 July 2026

Available Online: 16 August 2026

Revised: 8 August 2026

DOI: 10.5861/ijrse.2026.26368

Accepted: 15 August 2026

Abstract

This study examined the agreement between ChatGPT-4 Plus and experienced human raters in the assessment of argumentative essays written by Greek upper secondary school students. Specifically, it investigated (a) the degree of agreement between ChatGPT and human raters and (b) the relative contribution of content, text organization, and language (vocabulary and grammar) to overall writing scores. The dataset comprised 267 argumentative essays that had previously been evaluated by trained educators using the official Greek Ministry of Education scoring rubric. The same essays were subsequently assessed by ChatGPT-4 Plus using standardized prompting procedures. Results indicated moderate to good agreement between ChatGPT and human raters for both overall and trait-level scores, with agreement comparable to that observed between the two human raters. No significant differences were found in overall scores. At the trait level, however, ChatGPT assigned significantly higher scores for content, whereas human raters assigned higher scores for language; no significant differences emerged for text organization. Regression analyses showed that all three scoring components contributed significantly and relatively evenly to overall scores in both scoring systems, although ChatGPT demonstrated stronger intercorrelations among the scoring dimensions. Overall, the findings suggest that ChatGPT-4 Plus shows considerable promise as an automated essay scoring tool in Greek secondary education. However, systematic differences in evaluative emphasis indicate that it is currently best viewed as a complementary tool rather than a replacement for expert human judgment, particularly in high-stakes writing assessment.

Keywords: automated essay scoring, assessment tool, human raters, ChatGPT, Greek Upper Secondary Education

Evaluating ChatGPT-4 as an automated essay scoring tool: A comparative study with human raters in Greek upper secondary education

1. Introduction

The assessment of student writing has long been a cornerstone of language education and language testing, with implications for academic placement, curriculum development, and individual learner trajectories. Even so, the process of grading written assignments can be frustrating and overwhelming for teachers (Tsai, 2012). In recent years, the advent of Automated Essay Scoring (AES), Automated Essay Evaluation (AEE) and Automated Essay Grading (AEG) systems have introduced new paradigms for evaluating written language, offering the promise of efficiency, consistency, and scalability, although still controversial in several respects, but ultimately promising (Hockly, 2019). After all, the struggle for an “intelligent” system that can assist humans in education has a long history (Cope et al., 2021). The Automated Writing Evaluation (AWE) programs ultimately apply psychometric analyses to writing evaluation (Warschauer & Ware, 2006).

Among these systems, the emergence of large language models (LLMs), such as OpenAI's ChatGPT, has generated renewed interest and scrutiny regarding their potential role in educational assessment. After all, the use of AI is like utilizing hundreds if not thousands of human assessors to work in unison (Richardson & Clesham, 2021). That was a major breakthrough, since former AWE systems relied on quantitative computing methods for evaluation, overlooking the qualitative aspects since it was difficult to measure effectively in real life writing tasks (Huawei & Aryadoust, 2023). As a matter of fact, OpenAI's GPT -3, introduced in 2020, was a Natural Language Processing System (NLP) trained on a large corpus of text, being supervised and labelled, a process known as “fine-tuning” (Rudolph et al., 2023). This massive deep-learning language model contains approximately 1.3 billion parameters (Anson & Straume, 2022). These Neural Nets are multilayered statistical sequences for machine learning with human-like results (Cope et al., 2021). Even so, surprisingly enough, in the functionalities described in Anson and Straume (2022), the role of ChatGPT as an assessment tool for student essays is not included.

Nevertheless, there is still limited research on the topic of the implementation of ChatGPT in education, especially in assessment (Rudolph et al., 2023). Focusing on the use and implications of this topic is critical, as research suggests that the rapid advancement of AI is likely to transform education and human involvement in many tasks, potentially surpassing human performance—even in domains such as high school essay writing (estimated by 2026) (Grace et al., 2018).

This study examines the agreement between ChatGPT-4 Plus and experienced human raters in scoring authentic Greek upper secondary student essays using standardized national assessment criteria, providing evidence from a non-English, high-stakes educational context

2. Literature Review

2.1 Human vs Machine Writing Assessment®

Research comparing human and machine-based scoring of writing has proliferated in recent years, particularly in response to the increasing integration of AES tools in educational contexts (Koskinas et al., 2024; McCurry, 2010; Klobucar et al., 2013; Bridgeman & Ramineni, 2017). McCurry (2010) questioned whether machine scoring could handle open-ended, creative responses with the same interpretative flexibility as human readers, noting a performance gap in evaluating nuanced features like rhetorical strategy. Similarly, Klobucar et al. (2013) found that while AES systems achieved speed and consistency, they often lacked sensitivity to contextual factors influencing writing quality. Bridgeman and Ramineni (2017) emphasized the importance of aligning AES scoring

with authentic classroom practices to ensure contextual validity. Cope et al. (2021) proposed four transpositions for applying AI in education, namely: namability, calculability, measurability and representability to call this process reliable and trustworthy. In this validity argument Enright and Quinlan (2010) suggest evaluation, generalization, extrapolation and utilization. Finally, Koskinas et al. (2024) developed an AES (lexiscore®) taking into account the differential role of different categories to the assessment.

Although these studies collectively demonstrate significant advances in AES development, they also reveal an ongoing tension between computational efficiency and the complex, multidimensional nature of writing assessment. Most AES systems perform well when evaluating quantifiable linguistic features, yet they continue to struggle with higher-order dimensions such as creativity, audience awareness, critical reasoning, and communicative effectiveness. This suggests that improvements in scoring accuracy do not necessarily translate into improved construct representation, raising important questions regarding what AES systems actually measure and whether these measures fully reflect contemporary conceptions of writing competence.

More recent research has focused on the hybridization of scoring approaches. Enright and Quinlan (2010) proposed using AES systems to supplement, rather than replace, human raters in scoring essays by English language learners. Their findings suggested that e-rater® scores could enhance scoring reliability when used as a complement to human evaluation. Similarly, Chan et al. (2023) applied Rasch measurement to AES outputs, advocating for more nuanced, facet-based assessments. This hybrid approach was echoed by Attali et al. (2013), who proposed alternative procedures to enhance reliability through model-based calibration and iterative training of human raters. Furthermore, Ramineni and Williamson (2013) laid out psychometric guidelines for AES practices, reinforcing the necessity of transparency and validity in automated evaluation systems.

However, while hybrid assessment models are increasingly viewed as a practical compromise, relatively little consensus exists regarding the optimal balance between automated and human judgment. The literature frequently assumes that combining both approaches inherently improves assessment quality, yet empirical evidence regarding the contribution of each component across different educational contexts, genres, and learner populations remains limited. Consequently, further investigation is needed to determine under which conditions hybrid scoring provides meaningful advantages over either human or automated assessment alone.

2.2 *Validity and Reliability in AES Systems*

Validity and reliability remain the baseline of educational measurement (Bejar et al., 2014; Lee et al., 2009; Weigle, 2010). AES tools have been rigorously critiqued for potential construct underrepresentation and vulnerability to construct-irrelevant variance. Bejar et al. (2014) illustrated how AES systems could be misled by surface-level lexical or syntactic features, thus rewarding stylistic flourishes over substantive argumentation—a finding further supported by Vögelin et al. (2019).

Lee et al. (2009) developed a framework linking holistic and analytic scores with text features, revealing inconsistencies between human judgments and AES outputs. Similarly, Deane (2013) questioned the alignment of AES metrics with contemporary theories of writing, advocating for models that incorporate metacognitive, genre-based, and rhetorical constructs. Reiter and Belz (2009) contributed a technical evaluation of AES metrics' validity, raising concerns about model transparency and the explainability of AES decisions.

Collectively, these studies indicate that validity concerns extend beyond statistical agreement between AES and human raters. High correlations may obscure fundamental differences in the constructs being assessed, particularly if automated systems rely disproportionately on linguistic proxies rather than the communicative intent or argumentative quality of a text. Therefore, evidence of scoring consistency should not be interpreted as sufficient evidence of construct validity.

Douglas and Hegelheimer (2007) emphasized the significance of construct validity in computer-based assessment systems, noting that while automation enhances efficiency, it must not come at the expense of construct

coverage or interpretative fairness. Chapelle (1999) also addressed the foundational question of validity in language assessment, underscoring the need to align technological tools with credible assessment theory.

These theoretical perspectives remain particularly relevant in the era of AI-driven assessment, where increasingly sophisticated machine learning models often exhibit high predictive performance while providing limited interpretability. This creates a potential conflict between predictive accuracy and assessment transparency, especially in high-stakes educational settings where score explanations and fairness are essential components of validity.

Despite these concerns, AES systems have shown promising reliability metrics in controlled settings (Keller-Margulis et al., 2021). The work of Latifi and Gierl (2021) using Coh-Metrix features demonstrated successful AES applications in large-scale testing, although with caveats regarding genre and language level. In this regard Koskinas et al. (2024) found credible evidence for estimating the role of specific language features to overall assessment.

Nevertheless, the strong reliability reported in experimental and standardized contexts may not generalize to authentic classroom environments characterized by greater linguistic diversity, varied writing purposes, and heterogeneous learner populations. Consequently, future research should place greater emphasis on external validity, cross-context generalizability, and the ethical implications of AES implementation. Such work would help determine whether reliable automated scoring can also support equitable, pedagogically meaningful, and transparent assessment practices.

2.3 Writing assessment in the Greek Secondary Education Context

Writing assessment in the Greek secondary education system adheres to strict national rubrics, particularly in the upper levels (Koskinas et al., 2024). These rubrics are designed to ensure transparency, comparability, and fairness in high-stakes examinations, and they form the evaluative framework adopted in the present study. They comprise three interrelated dimensions that collectively capture argumentative quality, textual organization, and lexical sophistication.

More specifically, regarding the argumentative clarity, this criterion assesses the extent to which a student's central claim and supporting arguments are expressed in a precise, transparent, and logically accessible way. A text is considered strong in argumentative clarity if the thesis statement is explicitly identifiable, if the stance taken is not ambiguous, and if every argument contributes directly to the defense or development of that stance. In fact, argumentative writing is a high-stake task in standardized tests such as Scholastic Aptitude Test (SAT) since it demands structural, dialogical and language aspect of argumentation (Su et al., 2023).

On the other hand, coherence of structure evaluates how well the overall organization of the text contributes to its persuasiveness and readability. A coherent organization is marked by a logical progression of ideas from introduction to conclusion, with each paragraph serving a clear rhetorical function. Topic sentences, transitions, and signposting expressions play an essential role in maintaining this coherence, as they guide the reader through the argument.

Lexical accuracy focuses on the precision, appropriateness, and correctness of word choice. Lexical accuracy encompasses not only the avoidance of grammatical or vocabulary errors but also the ability to select terms that capture nuanced meaning and adhere to the register expected in academic writing. For instance, the use of discipline-specific terminology enhances credibility, while avoidance of vague or colloquial expressions ensures alignment with academic conventions.

1

The study underscores that evaluation cannot be divorced from cultural and educational contexts. Criteria developed in students' first language (L1) environments are shaped by local rhetorical traditions and expectations

of formality (Koskinas et al., 2024). Thus, what counts as “clear”, “coherent”, or “accurate” is partially determined by academic and cultural norms. Recognizing this interplay ensures that rubrics are applied not in a vacuum but with sensitivity to the broader frameworks that shape student writing. In addition, literature suggests that the relationship between the aforementioned parameters and the overall text quality is considered obscure (Lumley, 2002). The last remark should be taken into consideration, when we treat human assessment as norm or standard, even though the acceptance of human scoring is high (Enright & Quinlan, 2010). Zhang and Yuan (2022) demonstrated that machine learning approaches using Natural Language Processing (NLP) can identify compositional elements related to coherence and cohesion, which are often emphasized in writing rubrics. Nevertheless, this process seems to produce inadequate results on the overall assessment validity.

Furthermore, the high-stakes nature of writing evaluation in the Greek Secondary School, particularly in preparation for national university entrance exams, demands precision and fairness in scoring. The corpus of 267 texts used in this study, scored by veteran educators with national examination experience, provides a robust benchmark for comparative AES evaluation.

2.4 Theoretical Models of Writing Quality

Conceptualizing writing quality is foundational to both assessment design and interpretation (MacArthur et al., 2018; Read, 2023). After all, writing is social in its core and encompasses social purposes that writing directed solely toward a machine may violate this principle (Deane, 2013). Models of writing often encompass linguistic, cognitive, and sociocultural dimensions. MacArthur et al. (2018) identified linguistic predictors of writing quality, such as syntactic complexity and lexical diversity, as instrumental in evaluating developmental progress in student writing.

Read (2023) and Monteiro et al. (2023) have emphasized the growing sophistication in vocabulary assessment, particularly through the lens of semantic context indices and lexical sophistication. These findings are relevant for evaluating the vocabulary criterion within the Greek national rubric. Durrant and Brenchley (2023) explored the development of noun phrase complexity across genres in children’s writing, offering corpus-based insights that can inform AES calibration. Hwang and Kim (2023) proposed that syntactic complexity may serve as a predictor of writing proficiency, particularly for EFL learners.

Robinson (2001) examined how task complexity and production conditions influence linguistic output, a concern mirrored in AES contexts where prompt formulation may affect the quality and consistency of responses. Treffers-Daller et al. (2018) and Wingrove (2022) addressed lexical diversity and formulaic language, respectively, while Woods et al. (2023) advocated for a multi-measure approach to lexical diversity—highlighting its importance in evaluating argumentative writing.

Yu (2009), Mulder and Hulstijn (2011), and Nasser and Thompson (2021) further connected lexical diversity and syntactic complexity to CEFR levels and native speaker benchmarks. Zenker and Kyle (2021) cautioned that minimum text length must be considered when computing lexical indices, a factor of particular relevance to this study’s corpus. Albeit, the several processes implemented in mainstream AWE such as NLP or Latent Semantic Analysis (LSA) often present several limitations such as cultural backgrounds, the consideration of specific assessment purpose and -finally- the validity argument that mere statistical techniques fail to respond effectively (Elliot et al., 2013). Taking into account not only the number of parameters for an overall evaluation, but also monitoring the several interactions between them is a mountainous process with procedural complexities (Rotou & Rupp, 2020).

3. The Present Study

The literature reviewed reveals both the promise and the complexity of using automated systems to evaluate student writing. While AES tools have achieved notable success in large-scale testing scenarios, their alignment with construct-relevant traits, especially in linguistically and culturally specific contexts, remains an open question.

The integration of ChatGPT introduces further questions related to model transparency, adaptability, and sensitivity to rubric-based evaluation.

3.1 Research Questions

Given the Greek context, the high-stakes nature of writing assessment, and the availability of a well-documented, human-scored corpus, this study aims to address the following research questions:

1. To what extent does ChatGPT agree with experienced human raters in the evaluation of upper secondary students' written production using standardized national scoring criteria, and to what extent do the two differ in overall scores and trait-level components (content, organization, language)?
2. How do trait-level components (content, organization, language) contribute to overall writing scores, and to what extent do their relative weighting patterns differ between ChatGPT and human raters within the given culturally bounded educational context?

These questions seek to bridge the gap between technological potential and pedagogical reliability in writing assessment, with implications for the future integration of LLMs in educational evaluation.

3.2 Materials and Methods

A sample of 267 essays was collected from upper-secondary students (157 from second grade and 110 from third grade) enrolled in Greek public schools. The writing tasks were argumentative in nature and drawn from the Greek National Exam prompts, ensuring alignment with authentic educational assessment contexts (cf. Perelman, 2014; Shermis, 2014). All student data were anonymized in accordance with GDPR compliance protocols, which -we should note- are the same for both secondary and third grade essentially.

Each essay was independently evaluated by two experienced human raters, both certified educators trained in using the official Greek Ministry of Education rubrics. These rubrics include trait-specific components, content relevance and development, text organization, lexical and grammatical appropriateness, which served as the basis for both human and AI scoring (Zhang et al., 2022; Wilson et al., 2017). The mean grade from both raters is used as a final result.

3.2.2 Trait-Level Categorization

Trait-level dimensions -content, language (vocabulary/grammar), text organization-were coded and analyzed separately. Drawing from Zhang et al. (2022) and Aryadoust and Liu (2015), these components were operationalized as follows:

- Content: Relevance to the prompt, idea elaboration, and argument quality.
- Organization: Logical progression, paragraphing, coherence markers
- Language: Range, precision, and lexical sophistication

Furthermore, in alignment with the official guidelines of the Greek Ministry of Education, assessors are instructed to adopt the same trait-level dimensions when conducting evaluations. The detailed instructions governing this process together with the score corresponding to the student's performance, are set out below:

Content:

- 12-10 points: Excellent to very good response to all aspects of the topic. Excellent to very good, creative utilization of the main ideas and concepts of the reference text(s). Excellent to very good logical development of

ideas (coherence), within the word limit. Excellent to very good achievement of the communicative objective.

- 9-7 points: Very good to good response to all aspects of the topic. Very good to good creative exploitation of the main ideas and concepts of the reference text(s). Very good to good logical development of ideas (coherence), within the word limit. Very good to good achievement of the communicative objective.

- 6-4 points: Good to limited response to aspects of the topic. Good to limited exploitation of the main ideas and concepts of the reference text(s). Good to limited logical development of ideas (coherence), with slight exceeding of the word limit. Good to limited achievement of the communicative objective.

- 3-1 points: Limited to insufficient response to aspects of the topic. Limited to insufficient exploitation of the main ideas and concepts of the reference text(s). Limited to insufficient logical development of ideas (coherence), with greatly exceeding the word limit.

Text organization:

- 10-9 points: Excellent to very good text organization (coherence and logical progression of meanings, paragraphing and use of a wide variety of appropriate structural words). Excellent to very good utilization of the macrostructural features of the textual genre (e.g. address, diction, title, etc.). Excellent to very good ability to perform structural transformations to the reference text(s) (e.g. onomatopoeia, direct/indirect speech, density, etc.). Excellent to very good use of pragmatically appropriate style throughout the text.

- 8-7 points: Very good to good text organization (coherence and sequence of meanings, paragraphing, use of a small variety of appropriate structural words). Very good to good utilization of the macrostructural features of the textual genre (e.g. address, diction, title, etc.). Very good to good ability to perform structural transformations to the reference text(s) (e.g. onomatopoeia, direct/indirect speech, density, etc.). Very good to good use of pragmatically appropriate style throughout the text.

- 6-4 points: Good to limited text organization (coherence and sequence of meanings, paragraphing, use of limited variety or partially appropriate structural words). Good to limited utilization of the macrostructural features of the text genre (e.g. unsuccessful address, diction, title, etc.). Good to limited ability to structural transformations to the reference text(s) (e.g. unsuccessful nomenclature, unsuccessful conversion of direct/indirect speech, partial density, etc.). Good to limited use of pragmatically appropriate style (e.g. style mixing, pragmatically inappropriate style).

- 3-1 points: Limited to insufficient text organization (limited coherence and sequence of meanings, partial or total absence of paragraphing, absence of structural words). Limited to insufficient utilization of the macrostructural features of the textual genre (e.g. absence of address, diction, title, etc.). Limited to insufficient ability to structural transformations to the reference text(s) (e.g. absence of nomenclature, density, etc.). Limited to insufficient use of pragmatically appropriate style throughout the text (e.g. pragmatically inappropriate style).

Language (vocabulary/grammar):

- 8-7 points: Excellent to very good presence of a relatively broad of correct grammatical and syntactic structures, compatible with the communicative purpose, formulated with precision. Excellent to very good command of a wide range of appropriate lexical elements, compatible with the intended purpose and formulated with precision.

- 6-5 points: Very good to good presence of a relatively broad range of correct grammatical and syntactic structures, compatible with the communicative purpose, formulated with relative precision. Very good to good command of a wide range of relatively appropriate lexical elements, compatible with the intended purpose and formulated with relative precision.

- 4-3 points: Good to limited presence of correct grammatical and syntactic structures, compatible with the

intended purpose, formulated with precision. Good to limited presence of appropriate lexical elements, compatible with the intended purpose and formulated with precision.

- 2-1 points: Limited to insufficient presence of correct grammatical and syntactic structures, compatible with the intended purpose, formulated with precision. Limited to insufficient presence of appropriate lexical elements, compatible with the intended purpose and formulated with precision.

To evaluate ChatGPT performance on these traits, a comparative matrix was constructed matching AI scores against the mean human rater scores for each trait per essay.

ChatGPT Evaluation Protocol

The same 267 essays were input into OpenAI's ChatGPT-4 Plus, with clear prompts designed to simulate human rater behavior. This point is deemed crucial, since AI responds differently in differentiated feedback and is especially sensitive to prompt initial data, even in minor adjustments, in order to produce human-like responses (Barrot, 2023). During the initial stage of prompt design, the tool was assigned the role of evaluator to elicit more targeted and accurate outputs (Schulhoff et. Al, 2024). General guidelines were provided regarding the three parameters [content, language (vocabulary/grammar), text organization], followed by a detailed specification of the evaluation criteria associated with each parameter. These criteria were aligned with the official assessment rubric and served as the basis for determining the final scores assigned to each essay. For each essay, ChatGPT was instructed to generate both holistic and trait-level scores based on the official rubric descriptions. To reduce prompt variance and ensure replicability, all evaluations were conducted using a standardized template modeled after best practices in automated writing assessment (Guo et al., 2013; Shermis, 2014). The AI scoring sessions were conducted in a controlled environment with no external data access or contextual bias.

To ensure agreement and mitigate randomness in AI outputs, each essay was submitted two separate times using the same prompt, and the mean trait scores were recorded. This approach aligns with recent protocols in automated scoring research, emphasizing stability across runs (Vo et al., 2023).

The prompt was formulated in the Greek language, and its corresponding English translation is provided below:

“Assess the following written text, following the criteria applied by national examination context as an evaluator in national examinations would do.

In conducting the assessment, the following criteria should be taken into consideration: 1. Content, 2. Language of the written text 3. Text organization. Each criterion is to be evaluated separately. Do not take spelling errors into account when grading the written text. More specifically,

Content should be graded as follows: 12-10 points: Excellent to very good response to all aspects of the topic [...].

Language of the written text should be graded as follows: 8-7 points: Excellent to very good presence of a relatively broad of correct grammatical and syntactic structures, compatible with the communicative purpose, formulated with precision [...].

Text organization should be graded as follows: 10-9 points: Excellent to very good text organization (coherence and logical progression of meanings, paragraphing and use of a wide variety of appropriate structural words) [...].”

All the detailed instructions, as outlined in the previous subsection, were assigned to each trait-level.

3.3 Statistical Analysis

All analyses were conducted using IBM SPSS Statistics (version 30).

To address RQ1, which examined the degree of agreement between ChatGPT and human raters, inter-rater agreement was assessed using intraclass correlation coefficients (ICCs) based on a two-way mixed-effects model with absolute agreement and single measures. ICCs were calculated for overall scores as well as for each trait-level component [content, language (vocabulary/grammar), text organization]. In addition, paired-samples t-tests were conducted to compare mean differences between human raters and ChatGPT across overall and trait-level scores. Effect sizes were calculated using Cohen's d to assess the magnitude of observed differences.

To address RQ2, which investigated the contribution and relative weighting of trait-level components in overall scoring, Pearson correlation coefficients were computed to examine associations between each trait [content, language (vocabulary/grammar), text organization] and overall scores separately for human and ChatGPT ratings. Correlations among trait-level components were also examined to assess the degree of interrelatedness within each scoring system.

Subsequently, multiple linear regression analyses were conducted separately for human and ChatGPT ratings to examine the relative contribution of trait-level components to overall scores. In these models, overall score was treated as the dependent variable, while content, language (vocabulary/grammar), and text organization were entered simultaneously as contributors. Standardized regression coefficients (β) were used to compare the relative weighting of each component across rater groups.

Given the potential overlap among contributor variables, multicollinearity diagnostics were assessed using variance inflation factors (VIFs). Finally, because the overall score was derived from the same trait-level components within the scoring rubric, regression coefficients were interpreted as reflecting the relative weighting of components within each scoring system rather than independent prediction in a strict statistical sense.

3.4 Ethical Considerations

This research adheres to ethical guidelines for research involving AI and student writing. Students' identities were fully anonymized, and human raters were blinded to both student identities and AI scores in accordance with GDPR compliance protocols. All AI outputs were checked to ensure no content generation beyond evaluation occurred, in line with best practices for educational applications of large language models (Shermis, 2014).

4.1 Agreement between ChatGPT and human raters (RQ1)

To evaluate the agreement between ChatGPT and human raters, intraclass correlation coefficients (ICCs) were calculated. For the overall essay scores, agreement was moderate to good, $ICC = .67$, 95% CI [.59, .73], $p < .001$, indicating satisfactory consistency between ChatGPT and the human raters. At the trait level, agreement was comparable across the three scoring dimensions, with ICCs of .63 for content, .65 for text organization, and .65 for language (vocabulary and grammar), indicating moderate agreement across all traits.

For comparison, the agreement between the two human raters was $ICC = .65$, suggesting that ChatGPT achieved a level of agreement comparable to that observed between the human raters. In addition, ChatGPT demonstrated higher internal consistency across the three scoring dimensions ($ICC = .80$), indicating a more consistent application of the scoring criteria across traits.

Paired-samples t-tests were conducted to examine differences in mean scores between human raters and ChatGPT (see Table 1). Results indicated no statistically significant difference in overall scores ($t(266) = -1.29$, $p = .200$, $d = -0.08$). However, significant differences emerged at the trait level. ChatGPT assigned higher scores for content than human raters ($t(266) = -7.10$, $p < .001$, $d = -0.43$), whereas human raters assigned higher scores for language (vocabulary/grammar) ($t(266) = 3.11$, $p = .002$, $d = 0.19$). No significant difference was observed for text organization ($t(266) = 0.07$, $p = .946$, $d = 0.004$).

Table 1*Means (M), Standard Deviations (SD), and paired comparisons between human and ChatGPT scores.*

Variable	Human M (SD)	AI M (SD)	<i>t</i>	<i>p</i>	<i>d</i>
Content	69.5 (14.1)	75.1 (15.5)	-7.10	< .001	-0.43
Language (vocabulary/grammar),	69.4 (12.7)	66.6 (17.1)	3.11	.002	0.19
Text organization	71.8 (14.2)	71.8 (16.9)	0.07	.946	0.00
Total	70.3 (12.0)	71.2 (15.7)	-1.29	.200	-0.08

4.2 Contribution of trait-level components to overall scores (RQ2)

To examine the contribution of trait-level components to overall writing scores, Pearson correlations and multiple regression analyses were conducted separately for human and ChatGPT ratings. Pearson correlation analyses revealed strong positive associations between each trait-level component and overall writing scores for both human and ChatGPT ratings (see Table 2).

For human raters, correlations between trait-level components and overall scores ranged from $r = .87$ (text organization) to $r = .89$ (content), all $ps < .001$. Similarly, for ChatGPT, correlations ranged from $r = .87$ (text organization) to $r = .93$ (content), all $ps < .001$. Intercorrelations among trait-level components were moderate for human raters ($r = .63$ – $.70$) and notably higher for ChatGPT ratings ($r = .84$ – $.87$), suggesting that ChatGPT's trait-level evaluations are more strongly interrelated than those of human raters.

Table 1*Pearson correlations among trait-level components and overall scores for human and ChatGPT ratings (N = 267)*

Variable	1	2	3	4
Content	—	.87***	.86***	.93***
Language (vocabulary/ grammar)	.70***	—	.84***	.90***
Text organization	.65***	.63***	—	.87***
Overall	.89***	.88***	.87***	—

Note. Values above the diagonal represent ChatGPT ratings; values below the diagonal represent human ratings.

*** $p < .001$ (two-tailed)

To further examine the relative contribution of each scoring component, multiple linear regression analyses were conducted (see Table 3). For the human ratings, the regression model was statistically significant, $R^2 = .999$, $F(3, 263) = 114,982.37$, $p < .001$. All three components contributed significantly to the overall score: content ($\beta = .384$, $p < .001$), language (vocabulary and grammar; $\beta = .361$, $p < .001$), and text organization ($\beta = .392$, $p < .001$). Among the three predictors, text organization had the largest standardized regression coefficient, although the differences among predictors were small.

For the ChatGPT ratings, the regression model was also statistically significant, $R^2 = .999$, $F(3, 263) = 354,213.64$, $p < .001$. Again, all three components contributed significantly to the overall score: content ($\beta = .330$, $p < .001$), language (vocabulary and grammar; $\beta = .370$, $p < .001$), and text organization ($\beta = .351$, $p < .001$). Language (vocabulary and grammar) had the largest standardized regression coefficient, although the differences among predictors were modest. Multicollinearity diagnostics indicated acceptable levels for the human ratings (all VIFs < 2.50) and moderate but acceptable levels for the ChatGPT ratings (VIFs = 4.53–5.23), suggesting some overlap among the predictor variables but no evidence of problematic multicollinearity. Importantly, because the overall score is derived from the same trait-level components within the scoring rubric, these regression results should be interpreted as reflecting the relative weighting of components within each scoring system, rather than independent prediction.

Table 2*Multiple regression analyses examining the relative contribution of trait-level components to overall scores*

Human Ratings

Variable	<i>B</i>	<i>SE B</i>	β	<i>t</i>	<i>p</i>	VIF
Content	0.329	0.002	.384	148.78	< .001	2.30
Language (vocabulary/ grammar)	0.341	0.002	.361	143.25	< .001	2.19
Text organization	0.332	0.002	.392	165.77	< .001	1.93

Model fit. $R^2 = .9992$, $F(3, 263) = 114,982.37$, $p < .001$

ChatGPT Ratings

Variable	<i>B</i>	<i>SE B</i>	β	<i>t</i>	<i>p</i>	<i>VIF</i>
Content	0.335	0.002	.330	148.96	< .001	5.23
Language (vocabulary/ grammar)	0.340	0.002	.370	175.77	< .001	4.72
Text organization	0.325	0.002	.351	169.82	< .001	4.53

Notes. *N* = 267. Model fit, $R^2 = .9998$, $F(3, 263) = 354,213.64$, $p < .001$

Notes. VIF = Variance Inflation Factor. Values indicate acceptable multicollinearity for human ratings and moderate multicollinearity for ChatGPT ratings. Because the overall score is derived from the same trait-level components, regression coefficients reflect relative weighting rather than independent prediction.

The aim of this study was to examine the degree of agreement between ChatGPT-4, used as an AES system in the context of Greek secondary education, and experienced human raters in the evaluation of argumentative writing tasks. Specifically, we examined (a) the degree of agreement between ChatGPT and human ratings (RQ1) and (b) the relative contribution and weighting of trait-level components -content, language (vocabulary/grammar), text organization- in scoring (RQ2).

5.1 Agreement between ChatGPT and human raters (RQ1)

With regard to the first research question, the results indicate that ChatGPT demonstrates moderate agreement in scoring student writing. Agreement between ChatGPT and human raters was moderate to good at both the overall and trait levels, and importantly, this level of agreement was comparable to the agreement observed between human raters themselves. This finding is particularly significant, as it suggests that discrepancies between ChatGPT and human ratings fall within the range of variability typically observed in human scoring practices. Moreover, ChatGPT exhibited higher internal consistency across trait-level scores, indicating a more uniform application of scoring criteria. This aligns with prior research highlighting variability in human rating behavior (Honko et al., 2023) and suggests that AI-based systems may offer increased consistency in evaluation. At the same time, this consistency should be interpreted cautiously, as human raters may capture contextual, rhetorical, and stylistic nuances that automated systems may not fully represent.

Overall, these findings suggest that ChatGPT demonstrates moderate agreement as a scoring tool, achieving levels of agreement comparable to those observed between human raters. At the same time, systematic differences in trait-level scoring indicate that ChatGPT and human raters prioritize different aspects of writing quality, with ChatGPT emphasizing content and human raters placing relatively greater emphasis on language (vocabulary/grammar). These findings contribute to the broader AES literature by suggesting that recent advances in large language models may address some of the longstanding concerns regarding the reliability of automated scoring systems (McCurry, 2010; Klobucar et al., 2013). Compared to earlier AES systems, which often demonstrated weaker alignment with human judgments (Wang & Brown, 2007), ChatGPT appears to achieve substantially improved agreement levels, approaching those reported for established systems such as e-rater (Attali & Burstein, 2005; Enright & Quinlan, 2010). However, agreement alone does not imply equivalence in evaluative reasoning. Although overall scores did not differ significantly between ChatGPT and human raters, systematic differences emerged at the trait level. ChatGPT assigned higher scores for content, whereas human raters assigned higher scores for language (vocabulary/ grammar), with text organization showing comparable evaluations. This pattern suggests that, while ChatGPT aligns with human judgments at a global level, it prioritizes different aspects of writing quality.

5.2 Contribution of trait-level components to overall scores (RQ2)

With regard to the second research question, the results indicate that all three trait-level components -content, language (vocabulary/grammar), text organization- were strongly associated with overall scores for both human

and ChatGPT ratings. Importantly, regression analyses showed that these components contributed in a relatively balanced manner, with only minor differences in their relative weighting.

For human raters, text organization showed a slightly stronger contribution, whereas for ChatGPT, language (vocabulary/grammar) showed a marginally stronger contribution. However, these differences were small, suggesting that both human and AI raters rely on similar underlying dimensions of writing quality when forming holistic judgments. This convergence supports the interpretation that ChatGPT captures core aspects of argumentative writing competence, including content development and text organization. At the same time, the analyses revealed higher intercorrelations among trait-level components in ChatGPT ratings, as well as increased multicollinearity compared to human ratings. This pattern suggests that ChatGPT may apply a more unified or less differentiated construct of writing quality, whereas human raters appear to distinguish more clearly between components. In other words, ChatGPT's evaluations may reflect a more integrated representation of writing performance, while human scoring remains more analytically differentiated.

The findings suggest that both human raters and ChatGPT rely on all three components in forming overall judgments, with relatively balanced weighting across traits. However, slight differences emerged: human raters showed a marginally stronger emphasis on text organization, whereas ChatGPT placed slightly greater weight on language (vocabulary/ grammar). These differences, although small, suggest subtle divergences in how human and AI raters operationalize writing quality. These findings both align with and extend prior research. Earlier AES systems were often criticized for over-reliance on surface-level features such as lexical diversity or cohesion (Bejar et al., 2014; Vögelin et al., 2019). In contrast, the present results suggest that ChatGPT places substantial emphasis on semantic relevance and argumentative content, indicating a shift in construct representation. This shift may be attributed to the architecture of large language models, which are optimized to capture meaning, coherence, and contextual relationships across extended discourse. Nevertheless, ChatGPT's relatively lower emphasis on language (vocabulary/grammar) highlights an important limitation. Features such as rhetorical nuance, stylistic appropriateness, and register sensitivity remain areas where human raters retain an advantage. These aspects are inherently context-dependent and require socio-pragmatic judgment, which current AI systems may approximate but not fully replicate.

5.3 Implications for AES and educational practice

Taken together, the findings suggest that ChatGPT demonstrates promising agreement and construct alignment, while also revealing systematic differences in evaluative emphasis. From a theoretical perspective, this supports the view that modern LLM-based systems may overcome some limitations of earlier AES approaches, particularly in capturing meaning and argumentation. From an educational perspective, the results have important implications for education. ChatGPT may serve as a valuable complementary tool, supporting human raters by enhancing consistency and reducing workload. However, it should not be considered a replacement for human evaluation, particularly in tasks requiring sensitivity to rhetorical nuance and contextual appropriateness. A hybrid human–AI approach appears to be the most appropriate direction. Such an approach would combine the consistency and scalability of AI systems with the interpretive depth and contextual awareness of human raters. This aligns with prior research demonstrating the benefits of combining automated and human feedback in writing assessment (Sari & Han, 2024). At the same time, caution is warranted. Overreliance on AI tools may lead to unintended consequences, including reduced learner engagement or overdependence on automated feedback (Barrot, 2023). Therefore, the integration of ChatGPT into educational practice should be carefully designed to support, rather than replace, pedagogical processes.

5.4 Limitations and Future Research

Despite the encouraging findings, several limitations should be acknowledged. First, the study is based on a specific sample of Greek upper secondary students and argumentative writing tasks, which may limit the generalizability of the results. Future research should examine whether similar patterns of agreement and trait-

level weighting are observed across different learner populations, genres, and educational contexts. Second, the study relies exclusively on quantitative analyses, which do not fully capture how ChatGPT interprets complex features of writing such as creativity, rhetorical nuance, and stylistic appropriateness. Given the observed differences in trait-level scoring—particularly ChatGPT’s emphasis on content and lower sensitivity to language (vocabulary/ grammar)—qualitative analyses would provide valuable insight into the underlying evaluative processes of both human and AI raters. Third, the findings suggest that ChatGPT’s trait-level scores are more strongly interrelated than those of human raters, indicating a potentially more unified or less differentiated representation of writing quality. While this contributes to scoring consistency, it may also limit the model’s ability to distinguish between distinct dimensions of writing competence. Future studies should further investigate this issue, particularly through alternative modeling approaches or more fine-grained trait definitions. Fourth, issues of fairness and bias remain an important concern. As ChatGPT is trained on large-scale human-produced data, its evaluations may reflect embedded linguistic or cultural biases. Systematic fairness audits across diverse demographic and linguistic groups are therefore necessary to ensure equitable assessment practices. Finally, the study adopts a cross-sectional design, which does not capture how ChatGPT evaluates writing development over time. Longitudinal research is needed to examine the stability and sensitivity of AI-based scoring across different stages of learner development, as well as its potential role in formative feedback contexts.

6. Conclusion

In conclusion, the findings indicate that ChatGPT achieves levels of agreement with human raters comparable to human–human agreement, while also exhibiting distinct patterns in the weighting of writing components. These results suggest that ChatGPT functions as a relatively consistent essay-scoring system, but one that reflects a slightly different evaluative logic—one that emphasizes semantic content and coherence over stylistic nuance. As such, its role in writing assessment is best understood as complementary, contributing to more efficient and consistent evaluation while preserving the essential role of human judgment. More broadly, the study contributes to the growing literature on AI-assisted assessment by providing empirical evidence from a non-English, high-stakes educational context, highlighting both the potential and the current limitations of large language models in educational evaluation.

AI Use Disclosure. We used ChatGPT-4 to assess argumentative essays written by Greek upper secondary school students. All outputs were reviewed, verified, and edited by the authors. No confidential or personally identifiable data were entered into AI tools. The authors take full responsibility for the content.

Author contributions statement (CRediT). Koskinas Emmanouil: Conceptualization, Methodology, Data curation, Writing – original draft, Writing – review & editing, Chadjipapa Elina: Conceptualization, Methodology, Data curation, Writing – original draft, Writing – review & editing, Dosi Ifigeneia: Formal analysis, Writing – original draft, Writing – review & editing, Gavriilidou Zoe: Conceptualization, Writing – original draft, Writing – review & editing. All authors have read and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

7. References

- Anson, C. M., & Straume, I. (2022). Amazement and trepidation: Implications of AI-based natural language production for the teaching of writing. *Journal of Academic Writing*, 12(1), 1-9.
<http://dx.doi.org/10.18552/joaw.v12i1.820>
- Aryadoust V., & Liu S. (2015). Predicting EFL writing ability from levels of mental representation measured by Coh-Metrix: A structural equation modeling study. *Assessing Writing*, 24, 35-58.
<http://dx.doi.org/10.1016/j.asw.2015.03.001>
- Attali, Y., & Burstein, J. (2005). *Automated essay scoring with e-rater version 2.0 (ETS RR-04-45)*. Princeton, NJ: Educational Testing Service. <http://dx.doi.org/10.1002/j.2333-8504.2004.tb01972.x>
- Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, 30(1), 125–141.
<http://dx.doi.org/10.1177/0265532212452396>

- Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, 57, Article 100745. <http://dx.doi.org/10.1016/j.asw.2023.100745>
- Bejar, I., Flor, M., Futagi, Y., & Ramineni C. (2014). On the vulnerability of automated scoring to construct-irrelevant response strategies (CIRS): An illustration. *Assessing Writing*, 22, 48–59. <http://dx.doi.org/10.1016/j.asw.2014.06.001>
- Bridgeman, B., & Ramineni, C. (2017). Design and evaluation of automated writing evaluation models: Relationships with writing in naturalistic settings. *Assessing Writing*, 34, 62–71. <http://dx.doi.org/10.1016/j.asw.2017.10.001>
- Chan, K., Trevor Bond, T., & Yan, Z. (2023). Application of an Automated Essay Scoring engine to English writing assessment using Many Facet Rasch Measurement. *Language Testing*, 40(1), 61–85. <http://dx.doi.org/10.1177/02655322221076025>
- Chapelle, C. (1999). Validity in Language Assessment. *Annual Review of Applied Linguistics*, 19, 254–272. <http://dx.doi.org/10.1017/S0267190599190135>
- Chen, L., Zaharia, M., & Zou, J. (2023). How is ChatGPT's behavior changing over time? *arXiv*. 1(1), 1–23. <http://dx.doi.org/10.48550/arXiv.2307.09009>
- Cope, B., Kalantzis, M., & Searsmith, D. (2021). Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies. *Educational Philosophy and Theory*, 53(12), 1229–1245. <http://dx.doi.org/10.1080/00131857.2020.1728732>
- Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, 18, 7–24. <http://dx.doi.org/10.1016/j.asw.2012.10.002>
- Douglas, D., & Hegelheimer, V. (2007). Assessing Language Using Computer Technology. *Annual Review of Applied Linguistics*, 27, 115–132. <http://dx.doi.org/10.1017/S0267190508070062>
- Durrant, P., & Brenchley, M. (2023). Development of Noun Phrase Complexity Across Genres in Children's Writing. *Applied Linguistics*, 44(2), 239–264. <http://dx.doi.org/10.1093/applin/amac032>
- Elliot, N., A. Ruggles Gere, G. Gibson, C. Toth, C. Whithaus, & Presswood, A. (2013). Use and limitations of automated writing evaluation software. *WPA-CompPile Research Bibliographies*, 23. Retrieved [10 September 2018] from <http://comppile.org/wpa/bibliographies/Bib23/AutoWritingEvaluation.pdf>
- Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, 27(3), 317–334. <https://doi.org/10.1177/0265532210363144>
- Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). Viewpoint: When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research*, 729–754. <http://dx.doi.org/10.1613/jair.1.11222>
- Guo, L., Crossley, S., & McNamara, D. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, 18, 218–238. <http://dx.doi.org/10.1016/j.asw.2013.05.002>
- Hockly, N. (2019). Automated writing evaluation. *ELT Journal*, 73(1), 82–88. <http://dx.doi.org/10.1093/elt/ccy044>
- Honko, M., Neittaanmaki, R., Jarvis, S., & Huhta, A. (2023). Beyond literacy and competency – The effects of raters' perceived uncertainty on assessment of writing. *Assessing Writing*, 57, <http://dx.doi.org/10.1016/j.asw.2023.100768>
- Huawei, S., & Aryadoust, V. (2023). A systematic review of automated writing evaluation systems. *Education and Information Technologies*, 28(1), 771–795. <http://dx.doi.org/10.1007/s10639-022-11200-7>
- Hwang, H., & Hyunwoo, K. (2023). Automatic Analysis of Constructional Diversity as a Predictor of EFL Students' Writing Proficiency. *Applied Linguistics*, 44(1), 127–147, <http://dx.doi.org/10.1093/applin/amac046>
- Keller Margulis, M., Sterett, M., & Matta, M. (2021). Validity of automated text evaluation tools for written expression curriculum based measurement: a comparison study. *Reading and Writing*, 34, 2461–2480. <http://dx.doi.org/10.1007/s11145-021-10153-6>
- Klobucar, A., Norbert, E., Perry D., Oleksandr, R., & Kamal, J. (2013). Automated scoring in context: Rapid

- assessment for placed students. *Assessing Writing*, 18, 62–84.
<http://dx.doi.org/10.1016/j.asw.2012.10.001>
- Koskinas, E., Gavrilidou, Z., Andras, C. & Markos, A. (2024). Linguistic factors affecting L1 language evaluation in argumentative essays of students aged 16 to 18 attending secondary education in Greece, *Assessing Writing*, 60, 3-10, <https://doi.org/10.1016/j.asw.2024.100844>
- Latifi, S., & Gierl, M. (2021). Automated scoring of junior and senior high essays using Coh-Metrix features: Implications for large-scale language testing. *Language Testing*, 38(1), 62-85.
<http://dx.doi.org/10.1177/0265532220929918>
- Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, 19(3), 246-276. <http://dx.doi.org/10.1191/0265532202lt230oa>
- MacArthur, C., Jennings, A., & Philippakos, Z. (2018). Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction? *Reading and Writing*, 32(1), 1553–1574. <http://dx.doi.org/10.1007/s11145-018-9853-6>
- McCurry, D. (2010). Can machine scoring deal with broad and open writing tests as well as human readers? *Assessing Writing*, 15, 118-129. <http://dx.doi.org/10.1016/j.asw.2010.04.002>
- Mulder, K., & Hulstijn, J. (2011). Linguistic Skills of Adult Native Speakers, as a Function of Age and Level of Education. *Applied Linguistics*, 32(5), 475–494, <http://dx.doi.org/10.1093/applin/amr016>
- Nasseri, M., & Thompson, P. (2021). Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences. *Assessing Writing*, 47, Article 100511.
<http://dx.doi.org/10.1016/j.asw.2020.100511>
- Oka, R., Kusumi, T., & Utsumi, A. (2024). Performance evaluation of automated scoring for the descriptive similarity response task. *Scientific Reports*, 14, Article 6228. <http://dx.doi.org/10.1038/s41598-024-56743-6>
- Perelman, L. (2014). When “the state of the art” is counting words. *Assessing Writing*, 21, 104–111, <http://dx.doi.org/10.1016/j.asw.2014.05.001>
- Ramineni, C., & Williamson, D. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, 18, 25–39. <http://dx.doi.org/10.1016/j.asw.2012.10.004>
- Read, J. (2023). Towards a new sophistication in vocabulary assessment. *Language Testing*, 40(1), 40 –46.
<http://dx.doi.org/10.1177/02655322221125698>
- Reiter, E., & Belz, A. (2009). An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4), 529–558.
<https://doi.org/10.1162/coli.2009.35.4.35405>
- Richardson, M., & Clesham, R. (2021). Rise of the machines? The evolving role of AI technologies in high-stakes assessment. *London Review of Education*, 19(1), Article 9, 1–13. <http://dx.doi.org/10.14324/LRE.19.1.09>
- Robinson, P. (2001). Task Complexity, Task Difficulty, and Task Production: Exploring Interactions in a Componential Framework. *Applied Linguistics*, 22(1), 27–57. <http://dx.doi.org/10.1093/applin/22.1.27>
- Rotou, O., & Rupp, A. A. (2020). Evaluations of Automated Scoring Systems in Practice. *ETS Research Report Series*, 2020(1), 1–18. <http://dx.doi.org/10.1002/ets2.12293>
- Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, 6(1), 342–363.
<https://doi.org/10.37074/jalt.2023.6.1.9>
- Sari, E., & Han, T. (2024). The impact of automated writing evaluation on English as a foreign language learners' writing self-efficacy, self-regulation, anxiety, and performance. *Journal of Computed Assisted Learning*, 40, 2065-2080. <http://dx.doi.org/10.1111/jcal.13004>
- Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., & Resnik, Ps. (2024). The Prompt Report: A Systematic Survey of Prompting Techniques. preprint arXiv: 2406.06608. <http://dx.doi.org/10.48550/arXiv.2406.06608>
- Shermis, M. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a

- United States demonstration. *Assessing Writing*, 20, 53–76. <http://dx.doi.org/10.1016/j.asw.2013.04.001>
- Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing*, 57. <http://dx.doi.org/10.1016/j.asw.2023.100752>
- Treffers-Daller, J., Parslow, P., & Williams, S. (2018). Back to Basics: How Measures of Lexical Diversity Can Help Discriminate between CEFR Levels. *Applied Linguistics*, 39(3), 302–327. <http://dx.doi.org/10.1093/applin/amw009>
- Tsai, M. H. (2012). The Consistency Between Human Raters and an Automated Essay Scoring System in Grading High School Students' English Writing. *Action in Teacher Education*, 34(4), 328–335. <https://doi.org/10.1080/01626620.2012.717033>
- Vo, Y., Rickels, H., Welch, C., & Dunbar, S. (2023). Human scoring versus automated scoring for english learners in a statewide evidence-based writing assessment. *Assessing Writing*, 56. <http://dx.doi.org/10.1016/j.asw.2023.100719>
- Vögelin, C., Jansen, T., Keller, S., Machts, N., & Möller, J. (2019). The influence of lexical features on teacher judgements of ESL argumentative essays. *Assessing Writing*, 19, 50–63. <http://dx.doi.org/10.1016/j.asw.2018.12.003>
- Wang, J., & Brown, M. S. (2007). Automated Essay Scoring Versus Human Scoring: A Comparative Study. *The Journal of Technology, Learning and Assessment*, 6(2). Retrieved from <https://ejournals.bc.edu/index.php/jtla/article/view/1632>
- Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, 10(2), 157–180. <http://dx.doi.org/10.1191/1362168806lr190oa>
- Weigle, S., C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*, 27(3), 335–353. <http://dx.doi.org/10.1177/0265532210364406>
- Wilson, J., Roscoe, R., & Ahmed, Y. (2017). Automated formative writing assessment using a levels of language framework. *Assessing Writing*, 34, 16–36. <http://dx.doi.org/10.1016/j.asw.2017.08.002>
- Wingrove, P. (2022). Measuring the frequency of the academic formulas list across corpora: A case study based in TED talks and Yale lectures. *Applied Corpus Linguistics*, 2, Article 100012. <http://dx.doi.org/10.1016/j.acorp.2021.100012>
- Woods, K., Hashimoto, B., & Brown, E. (2023). A multi-measure approach for lexical diversity in writing assessments: Considerations in measurement and timing. *Assessing Writing*, 55, Article 100688. <http://dx.doi.org/10.1016/j.asw.2022.100688>
- Yu, G. (2009). Lexical Diversity in Writing and Speaking Task Performances. *Applied Linguistics*, 31(2), 236–259. <http://dx.doi.org/10.1093/applin/amp024>
- Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, 47, Article 100505. <http://dx.doi.org/10.1016/j.asw.2020.100505>
- Zhang, D., & Yuan, X. (2022). Intelligent Scoring of English Composition by Machine Learning from the Perspective of Natural Language Processing. *Mathematical Problems in Engineering*, Article 9070272. <http://dx.doi.org/10.1155/2022/9070272>
- Zhang, X., Lu, X., & Li, W. (2022). Beyond differences: Assessing effects of shared linguistic features on L2 writing quality of two genres. *Applied Linguistics*, 43(1), 168–195. <https://doi.org/10.1093/applin/amab007>