English Language & Literature Teaching, Vol.31, No.1 Spring 2025
https://doi.org/10.35828/etak.2025.31.1.27

Building an Automated Scoring System for EFL Learners' Paraphrases via a Customized GPT

Minkyung Kim
(Seoul National University)

Kim, Minkyung. (2025). Building an automated scoring system for EFL learners' paraphrases via a customized GPT. *English Language & Literature Teaching, 31*(1), 27-57.

This study investigates the potential of a customized GPT as an automated paraphrase scoring (APS) system for EFL learners' paraphrases with implications for reducing teachers' grading workloads and achieving unbiased rating in classroom settings. A total of 1,000 paraphrases written by 100 Korean EFL learners were evaluated with the analytic and holistic scoring rubrics. The analytic rubric included syntactic change, word change, semantic equivalency, and grammatical accuracy, which are crucial to paraphrasing. A mixed-methods approach was employed to evaluate the APS's reliability and effectiveness. Quantitative analysis examined the reliability and consistency of the APS were with Pearson and Intraclass Correlation Coefficients. Inter-rater reliability of the scores between APS and human raters was analyzed in various comparison and demonstrated strong alignment. Additionally, the consistency of APS across two rubrics indicated moderate reliability overall. Qualitative analysis further investigated the nature of the scores generated by the APS and its pedagogical implications. These findings suggest the APS via a custom GPT has its potential as an automated tool for writing assessment, providing reliable feedback to students while replacing human raters. Blending the automatic evaluation with customized GPTs into classrooms can be a solution for some challenges detected in manual scoring in educational context.

[paraphrasing/automatic scoring system/writing assessment/large language model]

I. INTRODUCTION

Paraphrasing involves modifying words and sentence structures while preserving the original meaning. In the field of English as a Foreign Language (EFL), it plays a crucial role

in enhancing various linguistic skills, particularly in reading and writing (Keck, 2006; McCarthy et al., 2009; Pecorari, 2003). In the EFL context, paraphrasing can enhance learners' comprehension of reading texts, promote critical thinking and fluency, and support the development of lexical and syntactic competence (Adams & Chuah, 2022; Ariyanti & Anam, 2021; Emran et al., 2024; Yamada, 2003).

Given its significance for EFL learners, paraphrasing is particularly essential in academic writing. Previous research has shown that EFL learners often struggle to appropriately incorporate source texts, which can lead to serious issues such as academic dishonesty and plagiarism (Ahn, 2022; Shi, 2004; Wang, 2016; Wette, 2010; Zhang & Yu, 2023). Paraphrasing serves as a practical solution to this challenge, enabling EFL learners to integrate external ideas ethically while maintaining originality—an issue that has gained increasing attention in recent years (Ahn, 2022; Choi, 2012; Choy & Lee, 2012; McDonough et al., 2014; Sung, 2011; Wette, 2010).

However, despite its importance of teaching paraphrasing, there is a lack of a large-scale test under a name of "a paraphrasing task" as well as widely accepted scoring criteria (Liu & Lin, 2022). This line of research usually developed their own paraphrasing tasks and rating rubrics for evaluation (Chen et al., 2015; Choi, 2012; Ji, 2018; Kim, 2020). There is no consensus for established scoring standards among scholars, which might lead to improper scoring with bias. Some studies preferred to utilize analytic scoring rubrics (Choi, 2012; Liu & Lin, 2022; Tran & Nguyen, 2022; Zhang, 2020) whereas others adopted holistic scoring methods (Chen et al., 2015; Ji, 2018, Soheim, 2017).

Most studies rely on a single scoring scale based on relative merits and context. However, using both holistic and analytic rubrics can enhance reliability. Kim (2020) employed this approach to assess the scoring reliability of paraphrasing tasks. Combining two rubrics offers advantages such as improved reliability, validity, balanced feedback, and comprehensive evaluation. However, managing both simultaneously can be burdensome for teachers and practitioners due to its time-consuming nature.

Another issue of the absence of an established scoring rubric in paraphrasing assessment is a significant challenge on the potential development of paraphrasing tests as large-scale tests in the future. This can be further exacerbated in countries such as Korea, where diverse writing performance assessments are not widely implemented in educational settings (Kim, 2020; Lee, 2010). With restricted testing materials in writing, paraphrasing tests can be a potential alternative with its high validity indicating sufficient chance to be modified and applied in forms of large-scale tests in the near future (Choi, 2012; Ji, 2018; Kim, 2020). Therefore, exploring methods to minimize bias and subjectivity while enhancing efficiency in paraphrase assessment is of great significance.

To address these gaps, automated essay scoring (AES) presents a viable solution. By offering consistent and unbiased evaluation, AES systems provide several advantages,

including high processing speed, cost-effectiveness, and time efficiency, while achieving reliability comparable to or even exceeding that of human raters (Bereiter, 2003; Chung & O'Neil, 1997; Dikli, 2006; Page, 1966; Susanti et al., 2023). However, it is hard to create a suitable or tailored model to be employed in certain context if we lack expertise in computer programming or natural language processing (NLP). For example, remodifying the existing AES models to use particularly for paraphrase assessment is inappropriate because specific scoring criteria are different between general essays and paraphrasing.

Recently, thanks to advancements in artificial intelligence (AI) and the emergence of large language models (LLMs) such as ChatGPT by OpenAI, automatic scoring for writing production becomes much easier than ever before (Holmes & Tuomi, 2022; Latif & Zahi, 2024; Son, 2024; Wu et al., 2023). Moreover, introduction of customizable GPT functions within ChatGPT in 2023 presented new opportunities for leveraging AI tools in education (Almasre, 2024), especially in the automated writing evaluation including the provision of feedback. Since the use of the customizing feature in ChatGPT is underexplored, little has been investigated on their application in automatically scoring paraphrased outputs for EFL learners.

Thus, the present study aims to address this gap by examining the feasibility of using a customized GPT as an automated scoring system for paraphrasing assessment. Specifically, the current study investigates the inter-rater reliability between human raters and the GPT tailored for automatic paraphrase scoring. This study also offers insights of integrating customized GPTs in the educational context while relieving burdens of teachers and educators in huge amount of essay scoring and helping create their own models with ease.

II. LITERATURE REVIEW

1. EFL Paraphrasing Assessment

1) Development of Paraphrasing Tasks

Studies in EFL writing have emphasized the significance of paraphrasing as a critical skill, particularly in academic contexts. Previous research has focused on pedagogical insights for teaching paraphrasing and the evaluation of the outputs produced by EFL learners (Chen et al., 2015; Choi, 2012; Ji, 2012; Ji, 2018; Kim, 2020; Sung, 2011; Wette, 2010).

Ji (2018) developed a paraphrasing task and collected data from 364 participants, ranging from middle school first graders to college freshmen in Korea. The findings revealed strong correlations between paraphrasing ability and TOEIC scores among undergraduates, while

correlations with self-assessment scores for secondary school students were moderate.[1] These results suggest that paraphrasing is a fundamental writing skill, which can reflect varying degrees of EFL learners' different levels of general English proficiency.

Other studies have primarily focused on paraphrasing tasks with undergraduate and graduate students, who might have been exposed more on employing paraphrasing skills in their school life through assignments and term papers. Chen et al. (2015) developed a corpus-based instructional program called PREFER which is aimed to help Chinese EFL learners improve their paraphrasing and general writing skills. They compared the paraphrased outputs of participants who used PREFER with those who relied on other tools during paraphrasing tasks.[2] The results demonstrated significant improvements in paraphrasing and overall writing skills among participants who used the program.

Kim (2020) recruited 100 Korean EFL learners in university and graduate school levels to take paraphrasing tests. The scores from the paraphrasing tasks demonstrated high inter-rater reliability between two expert human raters, as well as strong validity when compared to scores from official English tests such as TEPS, TOEFL, and TOEIC.[3] These results highlight the potential of paraphrasing tasks as reliable tools for assessing general English proficiency and writing skills of test-takers.

2) Rating Scales for Paraphrasing Tasks

All the paraphrasing tasks for EFL students were operating their own scoring scales regardless of the type of methods (e.g. analytic or holistic). As mentioned earlier, two possible choices of rating scales to be used in analyzing the reliability of paraphrasing tasks are analytic and holistic ones. Compared to paraphrasing assessment, in writing evaluation, it is well known that adopting holistic scales for grading is commonsense due to its absolute merit which is efficiency (Harsch & Martin, 2013). Following this trend, several research also chose holistic scoring method for assessing EFL paraphrases (Chen et al., 2015; Ji, 2018; Soheim, 2017).

For example, Chen et al. (2015) modified the rating scale from TOEFL Integrated Writing

---

[1] Ji (2018) adopted self-assessment scores for secondary school students due to the practical limitations of using standardized English proficiency tests, which are often unavailable or unsuitable for younger learners. These tests are primarily designed for higher proficiency levels and older students, such as university freshmen who typically take Test of English for International Communication (TOEIC) exams.

[2] Other tools included *Longman English Dictionary Online*, which provides single-word definitions, and *Thesaurus.com*, which suggests synonyms but lacks contextual information for effective paraphrasing.

[3] TEPS stands for Test of English Proficiency developed by Seoul National University. TOEFL stands for Test of English as a Foreign Language.

to develop their own holistic scoring rubrics which is structured with five levels (0-9 points) to assess the quality and quantity of students' paraphrasing performance. Similarly, Soheim (2017) explored the assessment of ESL students' paraphrasing skills through a structured test and a four-level holistic scoring rubric. The rubric evaluates paraphrasing based on accuracy, adherence to paraphrasing rules, and fluency.

However, these simple descriptions in the holistic scoring rubric can cause some misconceptions and low reliability at the end (Wiseman, 2012). To avoid this, others preferred analytic scoring scales for their paraphrasing tests because paraphrasing can be assessed with distinct sub-categories from general writings (Choi, 2012; Liu & Lin, 2022; Tran & Nguyen, 2022; Zhang, 2020). Liu and Lin (2022) employed an analytic rating scale consisting of three dimensions: (a) Restatement, (b) Meaning Consistency, and (c) Accuracy. They confirmed the analytic scoring rubric can provide detailed explanation on paraphrasing assessment in that it reflects the task construct. Tran and Nguyen (2022) utilized more sub-categories in their analytic scoring rubric: (a) Task fulfillment, (b) Structure, (c) Vocabulary, (d) Spelling, (e) Punctuation & Capitalization, and (f) Citation composed of five-level from 1 (lowest) to 5 (highest).

Research on holistic and analytic scoring has produced mixed findings regarding their impact on reliability and validity. Each method has its own advantages and limitations, making them suitable for different assessment purposes (Barkaoui, 2011). With this in mind, using two types of rubrics to be appropriately applied for the circumstances can be a potential solution. Kim (2020) employed both analytic and holistic scoring rubrics to examine score consistency in paraphrasing assessment. The findings indicated that, in comparison to the pilot study, which exclusively utilized a holistic scoring rubric, the combined use of both rubrics contributed to improved inter-rater reliability.

However, the issue of burden on manual scoring and chance of biased rating in large volume of materials have been reported in her study. The post-rating interviews in Kim (2020) revealed significant challenges in employing both holistic and analytic scoring rubrics and grading large volume of paraphrases. The raters noted that their adaptability to the rating process would have been improved if clear priorities within the analytic scoring criteria had been provided. This insight highlights the practical burden of human scoring of assessing huge amount of paraphrased responses, particularly when raters are tasked with applying complex scoring frameworks.

To address this problem, automatic essay scoring (AES) presents a potential solution for paraphrasing evaluation by enabling the automated grading of responses using both analytic and holistic scoring scales with high speed and efficiency. Thus, the current study focuses on applying AES to the assessment of paraphrased data from Korean EFL learners.

2. Automated Essay Scoring with Large Language Models

1) Brief History of Automatic Essay Scoring

Automated Essay Scoring (AES) uses technology to evaluate written work and assign numerical scores, providing an efficient alternative to traditional human scoring methods (Pack et al., 2024; Page, 1966; Ramesh & Sanampudi, 2022). Research on AES dates back to the late 20th century, with the development of diverse systems designed to enhance assessment efficiency and consistency. Among these, the *e-rater*, developed by Educational Testing Service (ETS), is one of the most widely recognized AES systems showing high reliability and validity in scoring essays (Attali & Burstein, 2004). High-stakes assessments, such as the Graduate Record Examination (GRE) and the TOEFL, have integrated AES systems to streamline the grading process and improve scoring efficiency while relieving bias and subjectivity.

Despite its potential, the use of AES has been met with some criticism (Condon, 2013; Pack et al., 2024; Perelman, 2014; Zhang, 2013). Zhang (2013) identified key limitations, including the lack of background knowledge in AES systems compared to human raters. Moreover, these tools may perpetuate biases and errors inherent in the human data used during their development.

However, some studies proposed automatic scoring introducing large language models (LLMs) can be a possible solution. According to Kasneci et al. (2023), LLMs like ChatGPT from OpenAI has its ability to provide precise, adaptive, and context-aware scoring by adding specific knowledge necessary in certain field. Latif and Zahi (2024) also argued that fine-tuning language models are emerging as possible alternatives especially in educational context. More details will be addressed in the next section.

2) Automatic Scoring with LLMs and Customization

Recent studies have demonstrated the potential of LLMs in assessing English learners' writing. Carlson et al. (2023) and Pack and Maloney (2023a) emphasized that LLMs not only provide reliable scoring but also offer deeper insights into students' writing. Unlike traditional automated essay scoring (AES) systems, which rely on predefined textual features, LLMs assess writing in a more flexible and context-sensitive manner (Ouyang et al., 2023).

Comparative studies further highlight GPT-4's superior performance in AES. Pack et al. (2024) found that GPT-4 outperformed Google's PaLM and Anthropic's Claude 2 in scoring 119 essays, achieving high reliability and validity in comparison to human ratings. Similarly, Mizumoto and Eguchi (2023) demonstrated ChatGPT's effectiveness in AES using 12,100 essays from the ETS TOEFL 11 corpus. However, these studies primarily focused on essay

scoring and did not explore customization to improve assessment reliability. While fine-tuning LLMs can enhance performance (Liu et al., 2023), it requires substantial computational resources and risks overfitting to training data. An alternative approach is the use of customized GPTs, introduced by OpenAI in November 2023, which allow users to tailor ChatGPT to specific domains without the complexity of fine-tuning (OpenAI, 2023b). Despite their potential, research on custom GPTs for automated assessment remains limited.

A notable exception is Almasre (2024), who applied a custom GPT to evaluate students' design projects in a typography course. The model demonstrated consistency in scoring, though differences in qualitative feedback were observed. While this suggests that custom GPTs can enhance assessment fairness and reduce human bias, no studies have systematically examined their reliability in paraphrase assessment, particularly for EFL learners.

To address this gap, the present study investigates the reliability of a customized GPT-based Automatic Paraphrase Scoring (APS) system as an alternative to human rating. Unlike previous AES research, which predominantly employs holistic rubrics (Mansour et al., 2024), this study examines both analytic and holistic scoring to provide a more comprehensive evaluation of paraphrase quality. By directly comparing APS scores with human ratings, this study aims to assess whether a customized GPT can achieve reliable paraphrase scoring while reducing subjectivity and grading workload of teachers in educational settings.

Based on these rationales, this study aims to examine the reliability of paraphrasing scores for Korean EFL learners using a custom GPT-based APS created with ChatGPT-4. To address these issues, the following research questions are proposed:

1. Do the APS system and human raters exhibit similar scoring tendencies across tasks for both analytic and holistic evaluations?
2. What is the inter-rater reliability between APS (a customized GPT for paraphrase assessment) and human raters in total scores for both analytic and holistic scoring?
3. How consistent is APS in providing both analytic and holistic scores?
4. What are the potential applications of the APS in classroom settings?

III. METHOD

1. Participants

The participants were undergraduate and graduate students aged between their early twenties and mid-thirties, with an average age of 24.52 years. This age group is considered

to be suitable for paraphrasing tasks, as these learners are likely to have more academic maturity and linguistic proficiency required for paraphrasing practice than elementary and secondary school students. Ji (2018) supports this, noting a strong correlation between paraphrasing performance and English proficiency among college-level students, whereas secondary school students showed weaker correlations due to their limited exposure to paraphrasing skills.

2. Paraphrasing Tasks

The dataset for this study consists of 1,000 paraphrased sentences produced by 100 Korean EFL learners.[4] The paraphrasing tasks were adapted from Chen et al. (2015) and modified to align with the study's objectives. Most target sentences in the tasks contain conjunctive adverbs and verb-noun phrases to facilitate participants to paraphrase and avoid any possible ambiguity. The participants completed five paraphrasing tasks, each composed of two sentences (see Appendix A). The number of tasks was set at five to ensure time and cost efficiency, while still providing sufficient data to evaluate participants' paraphrasing ability.

3. Paraphrase Scoring

1) Scoring Rubrics

Two types of scoring rubrics were employed to evaluate the paraphrases produced by Korean EFL learners: analytic and holistic scoring rubrics (see Appendix B). Each rubric offers distinct advantages and limitations. The holistic rubric provides clear and straightforward instructions, allowing for a quick and efficient scoring process. In contrast, the analytic rubric enables raters to gain detailed insights into the specific constructs of paraphrasing by evaluating multiple dimensions. While holistic rubrics are commonly used in writing assessments due to their simplicity, analytic rubrics offer greater depth but may impose a heavier burden on raters.

The analytic rubric used in this study included four dimensions critical for paraphrasing: (a) Syntactic Change (SC), (b) Word Change (WC), (c) Semantic Equivalency (SE), and (d) Grammatical and Mechanical Errors (G&M).[5] Considering that the paraphrasing skill is

[4] The data were originally derived from the author's master thesis and expanded for additional analysis in this study.

[5] In this study, SC is defined in terms of syntactic complexity, a well-established measure for assessing EFL learners' writing proficiency (Baek, 2023; Crossley, 2020; Ortega, 2003). Accordingly, in evaluating test-takers' paraphrasing ability, it is assumed that SC serves as an indicator of the degree of syntactic complexity present in their responses.

included in the area of writing, the grammatical and mechanical accuracy was regarded important. Scores were assigned on a 6-point scale ranging from 0 to 5 for each dimension, with a maximum possible score of 20 points per task.[6] In contrast, holistic scoring provided a single, overall score out of 5 points for each task. To assess the utility of the custom GPT as a rating tool, both the analytic and holistic rubrics were applied.

2) A Customized GPT as an Evaluator

The current study selected the GPT-4o as an automatic evaluator for assessing paraphrasing given that GPT-4o provides its customizing function and it was assumed to have its high reliability among its own versions (Pack et al., 2024). For automatic paraphrase evaluation, the APS model was designed through a GPT builder in ChatGPT as one of *my GPT*s. Open AI provides a service of creating customized GPTs where files can be uploaded. Paraphrased sentences, analytic and holistic scoring rubrics were uploaded to create the APS.

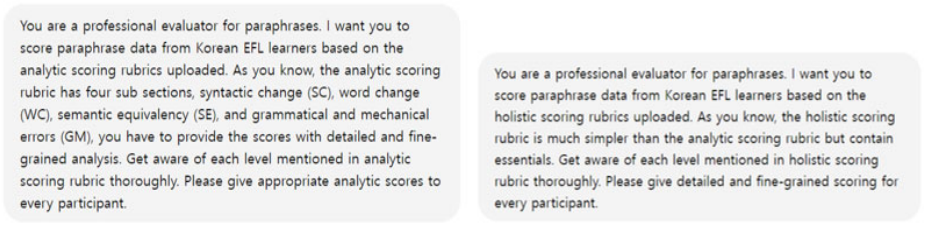
3) Automatic Scoring Procedure Using Customized GPT-4o

This study leveraged customizing feature of GPT-4o as an automatic evaluator for paraphrasing assessment, the APS. Due to the large volume of data from the Korean EFL learners paraphrase sentences, GPT-4o processed the ratings in batches. As the model could not process the entire dataset at once, the data were divided into smaller batches to ensure efficient and accurate evaluation.

The prompts designed for GPT-4o played a critical role in ensuring accurate and reliable scoring. Two distinct prompts were created to correspond to the analytic and holistic rubrics. These prompts provided detailed instructions to guide the model's evaluation process. The prompts used in this study are illustrated in Figure 1. As there were two kinds of scoring rubrics, analytic and holistic, the prompts are slightly different. For analytic scoring, the prompt directly mentions about the detailed criteria while the instructions for holistic scoring are rather simplified.

[6] Since this test was designed for EFL learners, grammatical and mechanical accuracy was considered equally important as syntactic and lexical modifications. As a writing performance measure, each dimension was assigned equal weight to ensure a balanced and comprehensive assessment of paraphrasing ability.

FIGURE 1
Prompts in the APS



A. The prompt given for analytic scoring

B. The Prompt given for holistic scoring

4) Human Raters

The original paraphrasing scores were assigned by two human raters with expertise in second language acquisition and English language teaching. To ensure consistency in scoring, the raters participated in training sessions conducted by the researcher. They were provided with two types of scoring rubrics—analytic and holistic—as well as a dataset consisting of paraphrased responses from 20 participants for pilot rating. Detailed scoring instructions were given for both rubrics, and the raters were required to adhere to these guidelines during the pilot rating process.

Following the pilot rating, the MAIN RATING SESSION was conducted for the remaining paraphrases. With 100 participants each paraphrasing 10 target sentences, a total of 1,000 paraphrases were evaluated. Each rater was responsible for assessing the entire dataset, including both the pilot and main scoring sessions. Additional adjustment sessions were conducted after both the pilot and main rating sessions to enhance scoring reliability. Discrepancies of more than two points between the raters' scores were reviewed, requiring raters to justify their ratings. For instances where the score differences exceeded three points, the raters engaged in discussions to explain their rationale and reach a consensus, thereby reducing scoring discrepancies. Inter-rater reliability was assessed using the Pearson correlation coefficient (PCC), yielding average correlation of .866 across all five paraphrasing tasks. This result indicates a high level of agreement between the two raters, demonstrating the reliability of the scoring process.

4. Data Analysis

For the current study, mixed method approach will be utilized. The quantitative investigation is directed to figuring out the rating tendencies among the different types of raters (the APS and human raters). Specifically, scoring patterns of the APS and human raters

will be analyzed through descriptive statistics and visualized in box plots. The mean, median, and standard deviation figures can show the similarities and differences among them.

To evaluate the inter-rater reliability between human raters and the APS in assessing paraphrasing data, the PCC was employed, a common metric in the AES research (Attali, 2013; Pack et al., 2024). PCC values were also calculated for both analytic and holistic rubrics to compare the reliability between two scoring rubrics. Additionally, individual dimensions within the analytic rubric were also calculated for reliability from SC to G&M.

To further assess the consistency of the model, the Intraclass correlation coefficient (ICC) was utilized. The ICC is widely recognized as a robust measure of agreement and consistency among raters (Mehta et al., 2018; Pack et al., 2024). The summary of variables and comparisons used in analyses were shown in Table 1. All statistical analyses were conducted using Python 3.1.3 to ensure accurate and reliable results.

In addition to the quantitative analysis, this study qualitatively examines the feasibility of the APS model by analyzing discrepancies between analytic scores assigned by human raters and APS. Sample rating has been presented to illustrate the potential advantages of APS in qualitative assessment. Furthermore, interviews with two human raters regarding their perceptions of the rating process are included in the results section.

TABLE 1
Summary of Variables, Comparisons, and Measures in Reliability Analyses

Comparison	Measures	Variables	Score Types
APS vs. Human Rater 1 and 2	PCC	APS analytic scores (total) Human rater analytic scores (total)	Total scores (analytic scoring)
		SC WC SE G&M	Individual scores (analytic dimensions)
		APS holistic scores Human rater holistic scores	Total scores (holistic scoring)
APS analytic vs. APS holistic scores	ICC	APS analytic scores APS holistic scores	Analytic vs. Holistic scoring

IV. RESULT

1. Paraphrasing Scores from the APS and Human Raters

1) Descriptive Statistics

This study aimed to assess the practicality of using a customized GPT model as an automatic scoring system for evaluating EFL learners' paraphrasing. The scores were obtained through two scoring rubrics of analytic and holistic scoring scales. Table 2 presents the descriptive statistics for the analytic scores across five tasks evaluated by the APS system, Rater 1, and Rater 2. For the APS system, Task 3 had the lowest mean score of 12.37 (*SD* = 4.63), while Task 4 recorded the highest mean score of 14.72 (*SD* = 5.75).

Similarly, for Rater 1, the scores ranged from a mean of 12.96 (*SD* = 5.07) for Task 1 to a mean of 14.60 (*SD* = 4.61) for Task 5. Rater 2's scores showed a similar pattern, with the mean scores ranging from 12.95 (*SD* = 4.82) for Task 1 to 14.05 (*SD* = 4.16) for Task 5. Across all tasks, the scores were fairly consistent between the APS system and both raters.

In holistic scoring, the APS system recorded the lowest mean score of 2.63 (*SD* = 1.25) for Task 1 and the highest mean score of 3.51 (*SD* = 1.10) for Task 5. For Rater 1, the mean scores ranged from 2.86 (*SD* = 1.29) in Task 1 to 3.42 (*SD* = 1.20) in Task 5. Rater 2 displayed lower scores overall, with the mean ranging from 2.48 (*SD* = 1.20) for Task 1 to 2.84 (*SD* = 1.23) for Task 3.

2) Grading Patterns

Figure 2 illustrates the grading patterns of the APS system and human raters through box plots. The score distributions for analytic scoring were largely similar, with mean scores ranging between 12.37 and 14.72. Rater 2 demonstrated lower variability in analytic scoring, as indicated by a narrower range between the maximum and minimum scores, whereas Rater 1 and the APS system exhibited greater variability. The close alignment between mean and median values suggests that score distributions were largely symmetrical. For holistic scores, notable differences emerged among the scoring methods. The APS system consistently assigned higher mean scores than both human raters, indicating a more lenient scoring tendency.

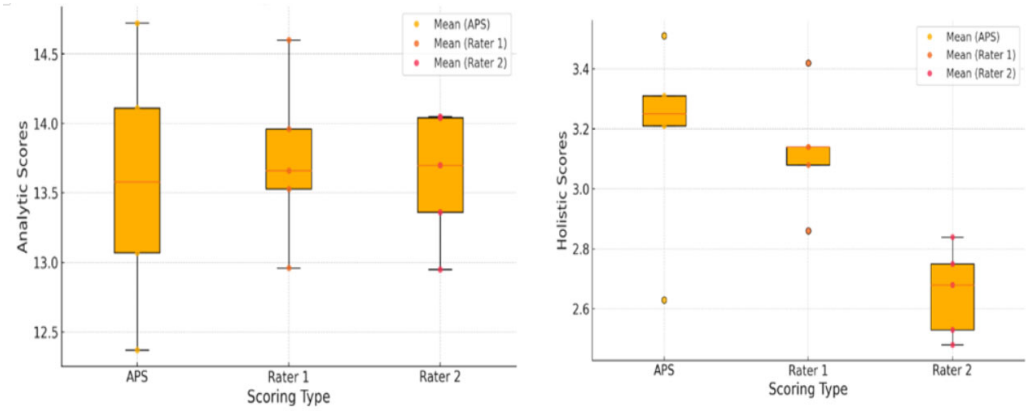
Rater 2, in contrast, produced the lowest mean scores but exhibited the greatest variability in holistic scoring, as reflected in its wider interquartile range. The APS system showed more uniform scoring patterns, whereas Rater 2 had the widest score distribution with greater spread and potential outliers, particularly in Task 1. Rater 1's holistic scores fell between those of the APS system and Rater 2, suggesting moderate alignment with both.

Overall, the findings suggest that the APS system provides more standardized and consistent scoring, particularly in holistic evaluations. While Rater 1's scoring aligned more closely with APS, Rater 2 demonstrated stricter judgments with greater variability in holistic scores. Despite differences in variability and central tendencies, the alignment of mean scores in analytic tasks suggests a level of agreement among scoring types, though human raters introduce more variation compared to the APS system.

TABLE 2
Descriptive Statistics of Analytic and Holistic Scores

			Analytic Scores				Holistic Scores		
Type	Task	*N*	Min	Max	Mean	*SD*	Max	Mean	*SD*
APS	Task 1	100	0	19	13.07	4.99	5	2.63	1.25
	Task 2			20	13.58	4.15		3.25	1.12
	Task 3			17	12.37	4.63		3.21	1.24
	Task 4			20	14.72	5.75		3.31	1.36
	Task 5			20	14.11	4.07		3.51	1.10
Rater 1	Task 1			20	12.96	5.07		2.86	1.29
	Task 2			20	13.66	4.48		3.08	1.17
	Task 3			20	13.96	4.88		3.14	1.37
	Task 4			20	13.53	5.62		3.14	1.44
	Task 5			20	14.60	4.61		3.42	1.20
Rater 2	Task 1			20	12.95	4.82		2.48	1.20
	Task 2			20	13.70	4.10		2.68	1.19
	Task 3			20	14.04	4.36		2.84	1.23
	Task 4			20	13.36	5.15		2.53	1.27
	Task 5			20	14.05	4.16		2.75	1.18

FIGURE 2
Differences in Scoring Patterns among the APS and Human Raters



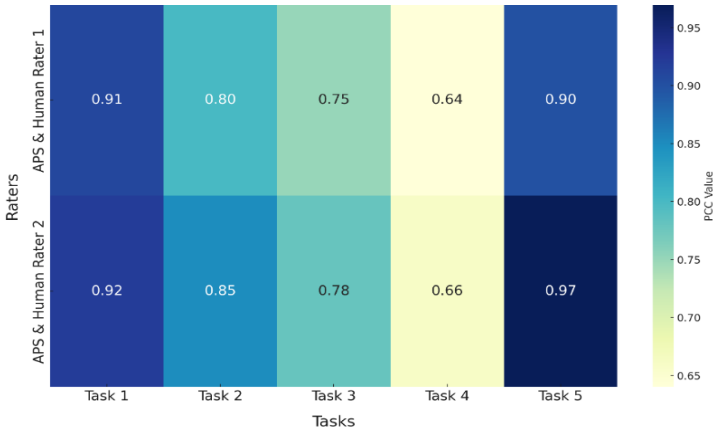
2. The Overall Inter-rater Reliability

The present study aimed to evaluate the feasibility of a customized GPT as an automatic paraphrasing scoring system for EFL learners' paraphrases. The first research question addresses this by investigating inter-rater reliability using Pearson Correlation Coefficient (PCC) values in comparing the total scores rated by the APS and human raters through analytic scoring rubric as shown in Figure 3. The rows correspond to the comparisons — APS with Human Rater 1 and APS with Human Rater 2 — while the columns represent the five paraphrasing tasks. The color intensity reflects the PCC values, with darker shades

indicating higher levels of reliability. Notably, all correlations were statistically significant at the .01 level.

The analysis reveals that Tasks 1 and 5 demonstrated excellent inter-rater reliability, with PCC values exceeding 0.90. Tasks 2 and 3 exhibited high reliability, with PCC values ranging between 0.75 and 0.85. In contrast, Task 4 displayed moderate reliability, with PCC values of 0.64 and 0.66 for APS with Human Rater 1 and APS with Human Rater 2, respectively.

FIGURE 3
Heatmap of Inter-rater Reliability between Human Raters and the APS



3. Reliability by Scoring Rubrics

The current study utilized two types of scoring rubrics—analytic and holistic—taking into account the potential needs and convenience of future users. The objective was to ensure that the scores generated by the APS system embodied via a customized GPT using these rubrics achieved high inter-rater reliability with human raters, thereby enhancing the system's applicability. To this end, inter-rater reliability was assessed for both individual scoring dimensions of the analytic scoring rubric (e.g. SC, WC, SE, and GM) and the overall scores from the holistic rubric.

Table 3 presents the inter-rater reliability values between the APS and Human Rater 1 based on the sub scoring criteria in analytic scoring rubric across all paraphrasing tasks. Among the five tasks, Task 1 exhibited the highest inter-rater reliability across all subcomponents except for grammatical and mechanical accuracy (GM). In contrast, Task 3 showed the lowest reliability in the syntactic change subcomponent, excluding GM as well. This lower reliability may be attributed to the characteristics of the original sentence in Task 3, which was notably shorter in length. While most Pearson Correlation Coefficient (PCC)

values indicated moderate reliability (ranging from 0.50 to 0.70), the lowest value in Task 3 was 0.48, slightly below the threshold for moderate reliability. Nonetheless, the majority of the subcomponents demonstrated moderate or higher reliability with Human Rater 1.

Table 4 displays the inter-rater reliability values between the APS and Human Rater 2 for the analytic scoring rubric. The results were generally similar to those for Rater 1 but showed slightly higher reliability overall. Notably, no PCC values fell below 0.50, indicating that all subcomponents achieved at least moderate inter-rater reliability with Rater 2. Additionally, the APS demonstrated excellent reliability (PCC values above 0.90) in four instances, further highlighting stronger agreement with Rater 2 compared to Rater 1.

The inter-rater reliability for scores rated using the holistic scoring rubric was also analyzed. Tables 5 and 6 present the PCC values for the APS compared to Human Rater 1 and Human Rater 2, respectively. Overall, the holistic scores achieved high inter-rater reliability across most tasks, with the exception of Task 2, where the reliability was only moderate. Interestingly, both human raters and the APS demonstrated moderate reliability for Task 2, suggesting a potential area for improvement in the APS assessment process. This finding may serve as a critical insight for refining the APS system in the future.

In sum, the APS achieved generally strong inter-rater reliability with both human raters across both analytic and holistic rubrics. However, certain tasks and individual scoring dimensions revealed areas for further enhancement. These findings provide valuable guidance for improving the APS's performance and its alignment with human evaluators.

TABLE 3
Inter-rater Reliability in Analytic Scoring Rubric with Rater 1

	Task 1	Task 2	Task 3	Task 4	Task 5
SC	.85	.60	.48	.52	.74
WC	.91	.63	.59	.85	.75
SE	.86	.66	.59	.62	.85
GM	.85	.71	.65	.64	.89

Note: All significant at .01 level (2-tailed)

TABLE 4
Inter-rater Reliability in Analytic Scoring Rubric with Rater 2

	Task 1	Task 2	Task 3	Task 4	Task 5
SC	.79	.66	.61	.55	.99
WC	.91	.77	.64	.77	.97
SE	.87	.80	.67	.69	.88
GM	.87	.69	.72	.64	.94

Note: All significant at .01 level (2-tailed)

TABLE 5
Inter-rater Reliability in Holistic Scoring Rubric with Rater 1

	Task 1	Task 2	Task 3	Task 4	Task 5
APS	.81	.61	.90	.92	.90

Note: All significant at .01 level (2-tailed)

TABLE 6
Inter-rater Reliability in Holistic Scoring Rubric with Rater 2

	Task 1	Task 2	Task 3	Task 4	Task 5
APS	.74	.54	.84	.82	.74

Note: All significant at .01 level (2-tailed)

4. Consistency of the APS

The third research question addresses the consistency of the APS model in evaluating paraphrased data. To investigate this, the ICC values were calculated to compare the APS ratings derived from the analytic and holistic scoring rubrics. ICC is a reliable measure for assessing consistency, indicating the degree of agreement between two or more rating methods.

Table 7 presents the ICC values for the five paraphrasing tasks. Four of the tasks—Tasks 1, 2, 3, and 5—achieved moderate reliability, with ICC values ranging from 0.50 to 0.63. These values suggest a reasonable level of agreement between the analytic and holistic scores for these tasks. Notably, Task 5 displayed the highest ICC value of 0.63, reflecting relatively stronger consistency between the two rubrics. In contrast, Task 4 exhibited an ICC value of 0.45, indicating weak consistency and highlighting discrepancies between the analytic and holistic scoring methods.

The overall results suggest that while the APS model generally demonstrates moderate consistency between analytic and holistic scoring, the variability in ICC values underscores the importance of task-specific factors in achieving alignment between the two scoring approaches (analytic and holistic scoring rubrics). Tasks with higher ICC values, such as Task 5, likely benefited from better-defined constructs that aligned well across rubrics, while tasks with lower ICC values, such as Task 4, may require revisions to scoring criteria or adjustments to the APS model for improved reliability. These findings provide valuable insights into the consistency of the APS model and emphasize the need for refining the scoring framework to minimize discrepancies between analytic and holistic rubrics.

TABLE 7
Intraclass Correlation Coefficient between Analytic and Holistic Scores of APS

	Task 1	Task 2	Task 3	Task 4	Task 5
APS	.58	.50	.55	.45	.63

Note: All significant at .01 level (2-tailed)

5. Qualitative Analysis

1) Rating Discrepancies

The present study examines qualitative aspects by analyzing differences in ratings between human raters and the APS model. Table 8 presents a sample paraphrase that received a score of 4 from the APS and a score of 2 from human raters for syntactic change, while for word change, the APS assigned a score of 3, and human raters assigned a score of 4. The scores for the remaining two dimensions were identical.

The APS model provided justifications for its assigned scores in the syntactic and word change dimensions. For syntactic change, the APS identified a clear structural transformation, noting the omission of "On the whole" and the replacement of "Similarly" with "Also," which altered the linking method. While the overall sentence order remained similar, the phrasing was sufficiently modified to reflect structural variation.

For word change, the APS recognized some lexical substitutions, such as the replacement of "risen" with "increased" and "increased quite considerably" with "risen a lot." However, it also noted that key terms, including "fuel prices" and "cost of food," remained unchanged, and "recent years" was directly repeated. Although some variation was present, the APS suggested that a greater range of word choices would improve the paraphrase.

These findings suggest that the APS system is capable of providing detailed, structured feedback on specific linguistic transformations, enabling a more transparent and consistent evaluation process compared to human raters. In large-scale assessments, human raters may not always explicitly articulate their reasoning or recall every rating with complete consistency, whereas the APS system ensures systematic and reproducible feedback across all responses.

TABLE 8
Sample Paraphrases with Distinct Scoring (the APS > Human Raters)

Participant #3	APS Rationale
Original Sentence On the whole, fuel prices have risen in recent years. Similarly, the cost of food has increased quite considerably. *Paraphrased Sentence* Fuel prices have increased in recent years. Also, the cost of food has risen a lot.	*Syntactic Change* The paraphrased sentences show a clear structural transformation. "On the whole" has been omitted, and "Similarly" has been replaced with "Also," which slightly alters the linking method. The overall sentence order remains similar, but the phrasing is sufficiently altered. *Word Change* Some words have been replaced, such as "risen" → "increased" and "increased quite considerably" → "risen a lot." However, "fuel prices" and "cost of food" remain unchanged, and "recent years" is directly repeated. While some variation exists, there could be more diverse word choices.

2) Human Rater Interview

The two human raters reported that they understood and followed the scoring rubrics well and were generally satisfied with the clarity of the analytic rubric. They found it sufficiently detailed to guide their evaluations, allowing them to complete the rating process without significant difficulty.

However, the raters faced challenges in holistic scoring, particularly in balancing a single overall score with the four analytic dimensions—syntactic change, word change, semantic equivalency, and grammatical and mechanical accuracy. Without explicit prioritization guidelines, they had to determine the weighting independently, leading to some uncertainty. They suggested that future training sessions provide clearer instructions on integrating analytic components into holistic scoring.

Additionally, they noted that scoring data from 100 test-takers was demanding, potentially contributing to rating fatigue. This underscores the limitations of manual scoring in large-scale assessments and highlights the potential of automated systems like APS to reduce workload while maintaining scoring consistency.

V. DISCUSSION

1. Reliability and Consistency of the APS

A primary purpose of this study was to evaluate the inter-rater reliability and consistency of the Automatic Paraphrase Scoring (APS) system, a customized GPT model designed to assess paraphrased sentences produced by EFL learners. By leveraging customization features, the APS was tailored to meet specific needs in grading paraphrases employing two kinds of scoring rubrics. The APS's performance was compared against human raters to examine its potential as a reliable and valid scoring tool.

The results of the current study demonstrated strong inter-rater reliability between the APS and human raters, as evidenced by significant positive correlations across both analytic and holistic scoring rubrics. The PCC values from total scores from analytic scoring ranged from 0.64, indicating moderate reliability, to 0.97, indicating near-perfect agreement. These findings suggest that the APS can consistently replicate human raters' evaluations, successfully serving a role as an alternative for human raters in the paraphrase assessment. Its ability to handle extensive datasets efficiently underscores its utility for reducing the burden of manual scoring in classrooms. For teachers and administrators managing high-stakes assessments, the APS represents a time-efficient replacement that ensures practicality without compromising reliability.

To examine the consistency of the APS, scores derived from analytic and holistic rubrics were compared using the ICC values. The analysis revealed moderate consistency across four of the five tasks, indicating reasonable alignment between the two scoring methods.[7] Among the paraphrasing tasks, Task 5 exhibited the highest consistency, with an ICC value of 0.63, reflecting strong agreement between analytic and holistic scores.

In contrast, Task 4 showed weak consistency, with an ICC value of 0.45, indicating notable discrepancies between the two scoring frameworks. These results suggest that while the APS generally maintains moderate consistency across scoring methods, certain task-specific factors, such as sentence length or complexity, may contribute to variations in reliability. The APS's ability to achieve reliability comparable to human raters reinforces its potential as a foundational component of automated writing evaluation systems.

[7] Reliability values in the range of 0.50–0.63 are considered moderate but acceptable for educational assessment research. According to *Shrout and Fleiss* (1979), ICC values between 0.40 and 0.75 indicate fair to good reliability, depending on the complexity of the assessment task.

2. Scoring Rubrics and Feedback

Using two different scoring methods of analytic and holistic rating was conducted to investigate the ability of the APS in terms of detecting the weaker spots in learners' paraphrasing performance and finally providing tailored feedback. The analytic rubric allows for detailed, section-specific feedback, making it highly beneficial for learners seeking to improve specific aspects of their paraphrasing skills. Conversely, the holistic rubric offers simplicity and convenience for graders, making it more suitable for large-scale assessments. Unlike most prior studies, which mainly utilized holistic rubrics in AES systems, the current study explored the potential of employing both types of rubrics within a single automated scoring framework.

When utilizing the analytic rubric, the APS demonstrated its ability to provide detailed feedback for each scoring dimension, as illustrated in Figure 4. By using prompts that specifically request explanations for assigned scores, the APS delivers clear and rubric-based justifications. Its ability to extract detailed feedback instantaneously represents a significant advancement over traditional feedback methods, which are often time-consuming and resource-intensive. Unlike human raters, who may struggle to maintain consistency and recall individual ratings in large-scale assessments, the APS system ensures systematic evaluation by providing structured and reproducible feedback across all responses.

This can benefit for both teachers and learners as well. In classroom settings, teachers can provide tailored and individual feedback for each student without too much burden. For learners who struggle to understand their performance and require further elaboration, the system can generate detailed explanations based on the uploaded scoring rubric. Providing directive feedback with explicit revisions and instructions has been shown to be effective in enhancing students' writing skills, particularly in grammar and mechanics (Kim & Joo, 2010). While engaging in this kind of interactive feedback process, they can improve paraphrasing skills on their own. Compared to human raters, the APS demonstrates qualitative advantages from test-takers' perspectives by providing detailed feedback while maintaining time and cost efficiency.

FIGURE 4
Detailed Feedback from APS

[Task 1]

Original Sentence: *On the whole, fuel prices have risen in recent years. Similarly, the cost of food has increased quite considerably.*

P035:

Overall, the oil prices have increased in the past few years. Simultaneously, the cost of living has increased extremely.

- **SC:** 4 (Good syntactic reorganization)
- **WC:** 4 (Good word changes with "cost of living" and "extremely increased")
- **SE:** 5 (Meaning is retained)
- **Grammatical Error & Mechanical Accuracy:** 4 (Minor awkward phrasing)
- **Total:** 17/20

3. Potential Improvements for Classroom Use

The findings of this study underscore the significant potential of a ChatGPT-based customized model for paraphrase assessment in classroom applications. By demonstrating reliability comparable to that of human raters and offering detailed feedback options, the Automatic Paraphrase Scoring (APS) system has the capacity to serve as a valuable tool for educators in facilitating effective language learning and assessment.

Previous research on paraphrase evaluation has largely focused on undergraduate and graduate students, often within academic or research-intensive contexts. However, the APS introduces new possibilities for expanding paraphrasing activities to a broader range of learners, providing a viable solution to the challenges associated with assessing writing performance in resource-limited settings. For example, simple paraphrasing tasks consisting of a few sentences could be incorporated into digital textbooks as interactive writing exercises for younger or less proficient learners. Teachers could instruct students to paraphrase newly acquired content, either as a performance assessment or as a routine homework assignment, with the APS facilitating the scoring process more efficiently than traditional methods.

When made accessible to both students and teachers, the APS has the potential to streamline the grading and feedback process, enhancing students' self-directed learning experiences while reducing the workload for teachers in evaluating daily paraphrasing tasks. By embedding paraphrasing activities into interactive learning materials, educators can foster foundational paraphrasing skills while utilizing the APS to deliver immediate and consistent feedback, ultimately contributing to more effective language instruction and assessment.

The APS's dual-rubric functionality further enhances its classroom utility by providing

flexibility for diverse teaching and assessment needs. Teachers can employ the analytic rubric to offer learners detailed, actionable feedback that supports skill development. Alternatively, the holistic rubric streamlines grading processes, making it particularly advantageous for large classes or high-stakes assessments. This adaptability ensures that the APS can effectively support both formative and summative assessments across varied educational contexts.

Moreover, the customization capabilities of GPT models significantly enhance the APS's practicality. Educators can align the system with specific curriculum goals or tailor it to address the unique needs of their learners, ensuring its relevance in diverse educational environments. By integrating APS-driven paraphrasing tasks into digital platforms, educators can not only reduce their workload and biased grading but also empower learners through immediate, personalized feedback that fosters skill development.

VI. CONCLUSION

This study explored the potential of a customized GPT model, the APS, for assessing Korean EFL learners' paraphrases, contributing to the emerging application of generative AI and LLMs in automatic essay scoring. Utilizing both analytic and holistic scoring rubrics, the APS demonstrated reliability in assessing paraphrased sentences in comparison to human raters. Although slight variations exist compared to human raters, the APS ensures consistency in its application of scoring criteria across all tasks, highlighting its flexibility and practical advantage for large-scale assessments. These findings hold implications for future research on developing automated evaluation procedures in other domains.

However, this study identified a few limitations inherent to optimizing and implementing the APS for large-scale paraphrase assessment. First, the process of configuring the system to assess all five paraphrasing tasks for 100 test-takers revealed challenges in the system's capacity to handle extensive datasets in a single session. Performance issues necessitated adjustments, such as dividing tasks into smaller batches and processing them incrementally, to ensure reliable operation. These findings highlight the broader need for further research into designing adaptable prompts and enhancing the capacity of AI-driven scoring systems to process large volumes of data effectively.

Additionally, differences in scoring criteria and task difficulty can be further analyzed using Multi-Facet Rasch Measurement (MFRM), which provides insights into rater characteristics, including severity, leniency, and task difficulty. Future research could enhance clarity by incorporating this method to refine the evaluation process.

Despite these challenges, the APS demonstrates strong potential as an automatic paraphrase scoring system. It achieves high correlations with human raters and offers the

flexibility to utilize both analytic and holistic scoring rubrics. This dual functionality makes it particularly well-suited for classroom settings, where it can provide tailored feedback to students with varying levels of paraphrasing ability. By supporting differentiated learning, the APS enables teachers to address diverse learner needs effectively and efficiently.

The findings also underscore the APS's broader potential for classroom application in the future. For instance, the APS can be integrated into digital English textbooks as a tool for assessing paraphrasing tasks in real time. Its ability to provide detailed and consistent feedback supports both formative and summative assessments, helping learners refine their paraphrasing skills while reducing the grading burden on teachers. Additionally, the APS's capacity to handle large datasets efficiently makes it suitable for large-scale classroom implementation, offering solutions for assessing paraphrasing activities.

In conclusion, while this study revealed important insights into the reliability and applicability of the APS, it also highlights areas for improvement that can guide future research. By addressing the identified challenges, researchers can further enhance the utility of automated scoring systems, expanding their role in advancing language education and assessment practices. Moreover, the integration of APS in classrooms provides opportunities for fostering language skills in EFL learners, bridging the gap between traditional assessments and innovative, AI-driven evaluation methods.

REFERENCES

Adams, D., & Chuah, K. M. (2022). Artificial intelligence-based tools in research writing: Current trends and future potentials. In P. P. Churi, S. Joshi, M. Elhoseny, & A. Omrane (Eds.) *Artificial intelligence in higher education* (pp. 169-184). CRC Press. https://doi.org/10.1201/9781003184157-9

Ahn, S. (2022). The effect of paraphrasing instruction on Korean EFL learners' attempts of paraphrasing. *The Journal of AsiaTEFL, 19*(2), 637-646. https://doi.org/10.18823/asiatefl.2022.19.2.16.637

Almasre, M. (2024). Development and evaluation of a custom GPT for the assessment of students' designs in a typography course. *Education Sciences, 14*(2), 148. https://doi.org/10.3390/educsci14020148

Ariyanti, A. & Anam, S. (2021). Technology-enhanced paraphrasing tool to improve EFL students' writing achievement and enjoyment. *JELTL (Journal of English Language Teaching and Linguistics), 6*(3). https://doi.org/10.21462/jeltl.v6i3.698

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 181-198). Routledge.

Attali, Y., & Burstein, J. (2004). Automated essay scoring with e-rater® v.2.0. *ETS Research Report Series, 2004*(2), i-21. https://doi.org/10.1002/j.2333-8504.2004.tb01972.x

Baek, S. (2023). A Corpus-based study of Korean EFL learners' use of subordinate conjunctions in argumentative writing: With focus on if, when, and because. *Journal of British & American Studies*, *58*, 145-170. 10.25093/ibas.2023.58.145

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing, 28*(1), 51-75. https://doi.org/10.1177/0265532210376379

Bereiter, C. (2003). Foreword. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary approach* (pp. vii–ix). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410606860

Carlson, M., Pack, A., & Escalante, J. (2023). Utilizing OpenAI's GPT-4 for written feedback. *TESOL Journal*, *15*, e759. https://doi.org/10.1002/tesj.759

Chen, M.-H., Huang, S.-T., Chang, J. S., & Liou, H.-C. (2015). Developing a corpus-based paraphrase tool to improve EFL learners'writing skills. *Computer Assisted Language Learning*, *28*(1), 22-40. https://doi.org/10.1080/09588221.2013.783873

Choi, Y. (2012). Paraphrase practices for using sources in L2 academic writing. *English Teaching*, *67*(2), 51-79. https://doi.org/10.15858/engtea.67.2.201207.51

Choy, S., & Lee, M. (2012). Effects of teaching paraphrasing skills to students learning summary writing in ESL. *Journal of Teaching and Learning, 8*(2), 77-89. https://doi.org/10.22329/jtl.v8i2.3145

Chung, K. W. K., & O'Neil, H. F. (1997). *Methodological approaches to online scoring of essays*. (ED418101). ERIC.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing, 18*(1), 100-108. https://doi.org/10.1016/j.asw.2012.11.001

Crossley, S. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, *11*(3), 415-445. https://doi.org/10.17239/jowr-2020.11.03.01

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5*(1). https://ejournals.bc.edu/index.php/jtla/article/view/1640

Emran, A. Q., Mohammed, M. N., Saeed, H., Abu Keir, M. Y., Alani, Z. N., & Ibrahim, F. M. (2024). Paraphrasing ChatGPT answers as a tool to enhance university students' academic writing skills. In *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems* (*ICETSIS*) (pp. 501-505). IEEE. https://doi.org/10.1109/ICETSIS61505.2024.10459386

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice, 20*(3), 281-307. https://doi.org/10.1080/0969594X.2012.742422

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education, 57*(4), 542-570. https://doi.org/10.1111/ejed.12533

Ji, N. (2012). Modes of paraphrasing attempted by Korean L2 writers. *English Teaching*, *67*(4), 131-148. https://doi.org/10.15858/engtea.67.4.201212.131

Ji, N. (2018). Investigation into validity of paraphrasing task as a writing performance test item for EFL learners. *Modern English Education*, *19*(2), 20-29. https://doi.org/10.18095/meeso.2018.19.2.03

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Keck, C. (2006). The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing*, *15*(4), 261-278. https://doi.org/10.1016/j.jslw.2006.09.006

Kim, M. (2020). *Investigating the reliability and validity of scores from a paraphrasing test for Korean EFL learners* [Unpublished master's thesis]. Seoul National University.

Kim, Y., & Joo, M. (2010). The effectiveness of diverse types of written feedback: Comparative study of teacher and student feedback. *English Language & Literature Teaching*, *16*(4), 133-152.

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence, 6*, 100210. https://doi.org/10.1016/j.caeai.2024.100210

Lee, W. (2010). Perceptions of high school students, their parents, and teachers on the factors determining the effects of the essay test on the curriculum of high schools in Korea. *KEDI Journal of Educational Policy, 7*(2), 209-232.

Liu, S., & Lin, D. (2022). Developing and validating an analytic rating scale for a paraphrase task. *Assessing Writing, 53*, 100646. https://doi.org/10.1016/j.asw.2022.100646

Liu, Z., He, X., Liu, L., Liu, T., & Zhai, X. (2023). Context matters: A strategy to pre-train language model for science education. *In International Conference on Artificial Intelligence in Education* (pp. 666-674). Springer Nature Switzerland.

Mansour, J., Patel, J., Singh, A., & Van Slyke, C. (2024). Panel: Generative artificial intelligence and information systems education. In *SAIS 2024 Proceedings*, 14. Association for Information Systems (AIS). https://aisel.aisnet.org/sais2024/14

McCarthy, P. M., Guess, R. H., & McNamara, D. S. (2009). The components of paraphrase evaluations. *Behavior Research Methods*, *41*(3), 682-690. https://doi.org/10.3758/BRM.41.3.682

McDonough, K., Crawford, W. J., & De Vleeschauwer, J. (2014). Summary writing in a Thai EFL university context. *Journal of Second Language Writing*, *24*, 20-32. https://doi.org/10.1016/j.jslw.2014.03.001

Mehta, S., Bastero-Caballero, R. F., Sun, Y., Zhu, R., Murphy, D. K., Hardas, B., & Koch, G. (2018). Performance of intraclass correlation coefficient (ICC) as a reliability index under various distributions in scale reliability studies. *Statistics in Medicine*, *37*(18), 2734-2752. https://doi.org/10.1002/sim.7679

MIT Sloan. (2024). Custom GPTs at MIT Sloan: A comprehensive guide. *MIT Sloan Educational Technology Office.* https://mitsloanedtech.mit.edu/ai/tools/writing/custom-gpts-at-mit-sloan-a-comprehensive-guide/

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

OpenAI. (2023a). *GPT-4* [Large language model]. https://openai.com/gpt-4

OpenAI. (2023b). Introducing GPTs. https://openai.com/blog/introducing-gpts

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*(4), 492-518. https://doi.org/10.1093/applin/24.4.492

Ouyang, S., Zhang, J. M., Harman, M., & Wang, M. (2023). *LLM is like a box of chocolates: The non-determinism of ChatGPT in code generation*. ArXiv. https://doi.org/10.48550/arXiv.2308.02828

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, *6*, 100234. https://doi.org/10.1016/j.caeai.2024.100234

Pack, A., & Maloney, J. (2023). Potential affordances of generative AI in language education: Demonstrations and an evaluative framework. *Teaching English with Technology*, *23*(2), 4-24. https://doi.org/10.56297/BUKA4060/VRRO1747

Page, E. B. (1966). The imminence of grading essays by computer. *The Phi Delta Kappan, 47*(5), 238-243. https://www.jstor.org/stable/20371545

Pecorari, D. (2003). Good and original: Plagiarism and patchwriting in academic second-language writing. *Journal of Second Language Writing*, *12*(4), 317-345. https://doi.org/10.1016/j.jslw.2003.08.004

Perelman, L. (2014). When 'the state of the art' is counting words. *Assessing Writing*, *21*, 104-111. https://doi.org/10.1016/j.asw.2014.05.001

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring system: A systematic literature review. *Artificial Intelligence Review, 55*, 2495-2527. https://doi.org/10.1007/s10462-021-10068-2

Shi, L. (2004). Textual borrowing in second-language writing. *Written Communication*, *21*(2), 171-200. https://doi.org/10.1177/0741088303262846

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420-428. https://doi.org/10.1037/0033-2909.86.2.420

Soheim, Y. (2017). Assessing ESL students' paraphrasing and note-taking. In S. Hidri & C. Coombe (Eds.), *Evaluation in foreign language education in the Middle East and North Africa* (pp. 169-183). Springer. https://doi.org/10.1007/978-3-319-43234-2_10

Son, H. (2023). Perceptions and attitudes of English learners towards the utilization of AI-based Automated Writing Evaluation (AWE) programs and instructor feedback. *English Language Teaching*, 29(3), 91-112. https://doi.org/10.35828/etak.2023.29.3.91

Sung, H. (2011). *Daehaksaeng yeongeo sseuygi-eyseo phyojeol yeybang-eul wuyhan inyoung-gwa hwaneonhagi hyullyeon hyogwa bunseok* [A study on the effects of quotation and paraphrase practices for avoidance of plagiarism for Korean university students] [Unpublished master's thesis]. Ewha Womans University.

Susanti, M. N. I., Ramadhan, A., & Warnars, H. L. H. S. (2023). Automatic essay exam scoring system: A systematic literature review. *Procedia Computer Science, 216*, 531-538. https://doi.org/10.1016/j.procs.2022.12.166

Tran, T. T. T., & Nguyen, H. B. (2022). The effects of paraphrasing on EFL students' academic writing. *Journal of Language and Linguistic Studies*, *18*(1), 976-987. https://doi.org/10.52462/jlls.233

Wang, W. (2016). Intertextual practices in academic writing by Chinese ESL students. *Applied Linguistics Review*, *7*(1), 53-72. https://doi.org/10.1515/applirev-2016-0003

Wette, R. (2010). Evaluating student learning in a university-level EAP unit on writing using sources. *Journal of Second Language Writing*, *19*(3),158-177. https://doi.org/10.1016/j.jslw.2010.06.002

Wiseman, C. S. (2012). A comparison of the performance of analytic vs. holistic scoring rubrics to assess L2 writing. *Iranian Journal of Language Testing, 2*(1), 59-92.

Wu, X., He, X., Liu, T., Liu, N., & Zhai, X. (2023). Matching exemplar as next sentence prediction (MeNSP): Zero-shot prompt learning for automatic scoring in science education. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova (Eds.), *Artificial intelligence in education. AIED 2023. Lecture notes in computer science* (Vol. 13916, pp. 389-400). Springer. https://doi.org/10.1007/978-3-031-36272-9_33

Yamada, K. (2003). What prevents ESL/EFL writers from avoiding plagiarism? Analyses of 10 North American college websites. *System, 31*, 247-258.

Zhang, D. (2020). An investigation of novice ESL writers' cognitive processes and strategy use of paraphrasing. *Language Testing in Asia, 10*(9), 1-26. https://doi.org/10.1186/s40468-020-00104-3

Zhang, D. E., & Yu, S. (2023). Investigating the relationship between linguistic changes in L2 writers' paraphrasing, paraphrasing performance, and L2 proficiency. *Applied Linguistics Review, 14*(5), 1451-1473. https://doi.org/10.1515/applirev-2021-0014

Zhang, M. (2013). Contrasting automated and human scoring of essays. *ETS R&D Connections, 21*(2), 1-11. https://www.ets.org/Media/Research/pdf/RD_Connections_21.pdf

APPENDIX A

An Example of a Paraphrasing Task

Paraphrasing Task

Please paraphrase the sentences given below. You can change the structure and vocabulary of the sentences but the meaning should be the same.

Example: A great number of people use Mandarin. Therefore, learning the Mandarin language has become a popular hobby. → Mandarin is spoken by a substantial number of people. Consequently, many people are interested in learning Mandarin.

1. Original Sentence: "Fuel prices have risen in recent years. Similarly, the cost of food has increased quite considerably."

2. Original Sentence: "The film drew attention to the fact that due to circumstances beyond our control, CO2 emissions could double by 2040. As a result, this could lead to worse natural disasters."

3. Original Sentence: "Laughter protects us from the damaging effects of stress. In other words, laughter is the best medicine."

4. Original Sentence: "Due to the weeklong rainfall beyond anyone's control, the flood destroyed many houses. As a result, many people became homeless."

5. Original Sentence: "There is no doubt that burning PVC plastic can severely damage human health. Therefore, some European countries have taken a lead in reducing the use of plastics."

APPENDIX B

Scoring Rubrics

1. Analytic Scoring Rubric

Level	Syntactic Change (Sentence Structure)	Word Change (Vocabulary/Expression)	Semantic Equivalency (Meaning)	Grammatical & Mechanical Accuracy
5	Significant syntactic changes in both sentences. Clear restructuring of sentence elements with improved readability.	More than three significant word or phrase changes. Changes show variety, precision, and appropriateness.	Precise retention of original meaning with no loss of information. The paraphrase is clear and easy to understand.	Flawless grammar, punctuation, and mechanics. No errors, and the sentence flows naturally.
4	One or two moderate syntactic changes in each sentence. Some parts remain similar, but there are noticeable improvements.	One to three significant word or phrase changes. Changes are generally accurate with minor reuse of original words.	The main meaning is preserved with minor shifts in nuance. The paraphrase remains generally clear.	Few grammatical or mechanical errors that do not affect comprehension. Sentence structure is generally fluent.
3	Minimal syntactic change. The paraphrase resembles the original with few attempts at restructuring.	One or two word changes that are mostly accurate but add little variety or improvement.	The core meaning is retained but some details or nuances are lost. Parts of the paraphrase may be confusing.	Some grammatical or mechanical errors that slightly hinder readability but do not obscure the meaning.
2	Little to no syntactic change. The structure is largely identical to the original with superficial adjustments.	Few or inaccurate word changes. The changes are minimal, redundant, or inappropriate.	The core meaning is altered or imprecise. Important details are missing or misrepresented.	Frequent grammatical or mechanical errors that significantly hinder readability and comprehension.
1	No syntactic change. The paraphrase uses the same structure as the original without any meaningful transformation.	No word changes or inappropriate substitutions. The paraphrase closely mirrors the original.	The meaning is significantly altered or lost. The paraphrase may misrepresent the original.	Numerous grammatical or mechanical errors that make the paraphrase hard to understand.

0	No attempt to paraphrase. The sentence structure is identical, or the participant did not follow directions.	No word changes or inappropriate substitutions that result in a nonsensical paraphrase.	The paraphrase is incomprehensible or irrelevant. The meaning is completely lost.	Grammatical and mechanical issues render the paraphrase incomprehensible. Severe errors throughout.

2. Holistic Scoring Rubric

Level	Criteria	Characteristics
Level 5 (Excellent Paraphrase)	- More than three word/phrase changes that appropriately replace the original wording. - Significant syntactic changes (e.g. changing sentence structure, passive to active). - Meaning perfectly preserved and easy to understand. - No grammatical or mechanical errors.	Sounds natural and well-constructed. Demonstrates a high level of linguistic sophistication.
Level 4 (Good Paraphrase)	- Two to three word/phrase changes that are appropriate. - Some syntactic changes, though not as complex (e.g., simple reordering of phrases). - Meaning clearly preserved with minor nuances. - Few or no grammatical or mechanical errors (minor mistakes acceptable).	Clear and understandable with some variation from the original. Demonstrates moderate linguistic competence.
Level 3 (Adequate Paraphrase)	- One to two word/phrase changes that are appropriate but do not significantly alter the sentence. - Minimal syntactic changes, closely resembling the original. - Meaning mostly preserved but harder to follow or slightly unclear. - Some grammatical or mechanical errors that do not hinder comprehension.	Retains much of the original sentence structure and wording. Errors do not distort the meaning.
Level 2 (Weak Paraphrase	- Few or no word changes, mostly using similar words to the original. - Little or no syntactic changes, very close to the original. - Meaning partially preserved, but some details are vague or incorrect. - Multiple grammatical or mechanical errors affecting clarity.	Demonstrates limited linguistic variation. Errors reduce the quality and make understanding difficult.
Level 1 (Poor Paraphrase)	- No word or phrase changes, or inappropriate changes that distort the meaning. - No significant syntactic changes, essentially a copy of the original sentence. - Meaning not preserved, or the paraphrase is confusing or incorrect. - Frequent grammatical and mechanical errors that significantly hinder comprehension.	Shows little effort to rephrase the original. Errors make understanding difficult or impossible.
Level 0 (No Paraphrase / Single Sentence)	- The participant did not follow directions (e.g., summarizing, writing a single sentence, or copying the original sentence without changes).	Fails to paraphrase in any meaningful way.

	- Meaning not preserved, or response does not fulfill task requirements.	

Examples in: English

Applicable Language: English

Applicable Levels: Tertiary

Minkyung Kim (student)
Dept. of English Language and Literature, Humanities Division
Seoul National University
08826, Gwanakro 1, Gwanak-gu, Seoul, Korea
Email: minkyung.kim0322@gmail.com

Received: January 31, 2025
Reviewed: March 4, 2025
Revised version received: March 25, 2025