Can ChatGPT Reliably and Accurately Apply a Rubric to L2 Writing Assessments? The Devil is in the Prompt(s)
(用 ChatGPT 评估二语写作的有效性研究：指令设计的重要作用)

Poole, Frederick J.
Michigan State University
(密西根州立大学)
poolefre@msu.edu

Coss, Matthew D.
Michigan State University
(密西根州立大学)
mattcoss@msu.edu

Abstract: This paper investigates the effectiveness of ChatGPT, a generative AI tool, in assessing second language (L2) writing. The study explores the practicality of employing ChatGPT as an assessment tool, focusing on the accuracy and reliability of the AI-generated scores compared to human raters. Various prompting strategies were tested to understand their impact on the effectiveness of ChatGPT in this context. The paper also examines the reliability of ChatGPT scores across different writing topics. The findings demonstrate that ChatGPT can serve as a valuable tool in L2 writing assessment, provided that it is used strategically with well-crafted prompts. The study contributes to the growing body of research on automated writing assessment tools, particularly in the realm of L2 learning, and offers insights into the practical application of such tools in educational settings.

摘要：本文研究了 ChatGPT 在评估二语（ L2 ）写作中的有效性。研究探讨了使用 ChatGPT 作为评估工具的实用性，重点关注 AI 生成的评分与人工评分相比的准确性和可靠性。为了理解不同指令策略对 ChatGPT 在这一背景下有效性的影响，本研究测试了十种指令策略。本文还检验了 ChatGPT 在不同写作主题上的评分可靠性。研究结果表明，ChatGPT 可以作为 L2 写作评估中的一个有价值的工具，前提是要使用精心设计的指令策略。本研究为自动写作评估工具的研究提供了新的见解，特别是在 L2 学习领域，并提供了这些工具在教育环境中实际应用的见解。

Keywords: Artificial intelligence; automated essay scoring; second language writing; writing assessment; rubrics

关键词: 人工智能，自动作文评分，二语写作，写作评估，评分标准

1. Introduction

In the last year, ChatGPT and other generative Artificial Intelligence (AI) tools have taken the world by storm. ChatGPT was one of the fastest platforms to reach 1 million users and has continued to experience sustained growth and use since its release in November of 2022. Since then, numerous tools with similar functionalities have emerged including Gemini (Google), Claude (Anthropic), and Perplexity (Perplexity.ai), among others. These generative AI tools, also referred to as large language models (LLMs), make use of recent developments in deep neural networks called transformers to optimize text generation capabilities (Vaswani et al., 2017). For many language teachers, administrators, and researchers, the introduction of generative AI tools like ChatGPT into the educational landscape is both exciting and intimidating. These tools have incredible capabilities, making them appealing for a variety of efficiency-improvement purposes. However, uncertainty due to the complexity of these tools' technological underpinnings as well as their trustworthiness for educational purposes remain strong as well. The last year has seen countless examples of users experimenting with ChatGPT and other AI tools to explore their capabilities and limitations. One area that may be particularly appealing to language educators is the potential of using tools like ChatGPT for assessment purposes. Evaluating writing assessments in the L2 classroom can be both time consuming and taxing (Crusan et al., 2016). Asking AI tools to apply a rubric to automatically evaluate student essays would undoubtedly sound attractive to many educators. However, before these tools can be normalized for assessment purposes, it is important to explore approaches to ensure high levels of reliability and accuracy while also considering these tools' practical relevance for teachers and other end-users given their inherent complexity. In other words, while AI tools like ChatGPT offer exciting prospects for optimization in education, the extent to which they are usable and useful to teachers (among others) must be thoroughly explored before recommendations can be made.

Many scholars have noted the time-consuming nature of evaluating assessments (e.g., Crusan et al., 2016) and the difficulty of avoiding human-rater bias and error (e.g., Elder et al., 2007). AI tools seem to be able to assess large amounts of data, including additional language (L2) learner writing, accurately and reliably (e.g., Mizumoto & Eguchi, 2023), but whether such tools can be implemented in the classroom in a practical manner remains unexplored. In this paper, we assert that the primary affordance of ChatGPT as an assessment tool lies in its capacity to expand the analytical capabilities of language educators, assessment specialists, and other professionals by making advanced computational techniques more accessible, regardless of the user's prior technical experience. Thus, we set out to explore strategies for prompting ChatGPT to produce reliable, accurate, and interpretable results for L2 writing assessments. We focus on prompting strategies, as we argue that this is the most accessible and impactful strategy for language educators to employ ChatGPT as an automated assessment scoring tool.

In this study, we analyze a series of prompts to demonstrate how effective prompting can empower teachers, our primary stakeholders, to employ ChatGPT successfully, bypassing some of the technical knowledge required to extract usable information from assessment data in prior research (see Mizumoto & Eguchi, 2023). Based on our analysis of these prompts and their different levels of reliability, we offer language educators and other language program stakeholders a list of considerations to improve reliability of generative AI as assessment scoring tools, with important emphasis on how and when these tools should or should not be used.

2. Literature review

2.1 Automated writing assessment tools

There is a long history of research on developing automated writing assessment tools. Much of this research explores tools created by large testing or publishing companies such as *e-rater* by Educational Testing Service, *Intelligent Essay Assessor* by Pearson Education, or *Intellimetric* by Vantage Learning among others (Hussein et al., 2019). These systems typically include both an automated scoring system as well as an automated feedback system. Research exploring automated feedback in these systems tends to focus on student and teacher perceptions of feedback and the impact of the tool on writing quality (e.g. Link et al., 2022). In contrast, research exploring the automation of assessment scores focuses on how similar automated scores are to human raters. In this study we are primarily concerned with automated scoring which is often referred to as automated essay scoring (AES).

AES systems have been used primarily in high-stakes assessments due to the cost of developing them. The most common approach to developing AES systems involves first using human raters to evaluate essays. Then collecting numerous automatically generated indices of text quality, and finally applying statistical approaches to identify which combination of these indices correlate with human scores best (Attali, 2015). Through the years these AES tools have advanced by adding more complex indicators such as readability scores and other text features extracted with natural language processing techniques (e.g., cohesion scores, syntactic complexity), as well as more complex statistical approaches (e.g., Bayesian text classification, Deep Neural Networks) (Huawei & Aryadoust, 2023; Hussein et al., 2019).

Several systematic reviews illustrate that AES tools can be quite accurate, but results vary substantially (e.g., Hussein et al., 2019; Ramesh & Sanampudi, 2022). While most studies have found that AES tools tend to correlate strongly with human scoring (>.7), some studies have noted inaccuracies. For instance, Wang and Brown (2007) found that over 25% of students received failing scores for a writing placement test (for L1 speakers) by human raters, while only 2% received a similar score by the AES tool. Wang (2015)

found that while EFL learners appreciated the quick feedback from an AES tool, *Criterion*, only 8% of students (n=53) who used the tool believed that it applied the writing rubric objectively and reliably to their writing. Furthermore, scholars have argued that AES tools may both misrepresent the writing construct and encourage a change in writing behavior to take advantage of weighted scoring systems (Deane, 2013). It is important to note that research on AES tools has been primarily (>90%) conducted with English language learners or L1 speakers of English (Huawei & Aryadoust, 2023), with few studies exploring other languages. Additionally, Reilly et al. (2014) noted in their study using an AES tool in an open online course that the AES tool was more accurate for L1 speakers of English than for L2 speakers of English. Qian et al., (2020) evaluated the *iWrite* system for L2 learners of English in China and concluded that the system failed to report accurate scores reliably. Thus, while these AES tools are continuing to improve, there is still some concern in terms of how accurately they are able to assess the written output of L2 learners.

Although much of the research has focused on the English language, there is a growing body of research on AES tools for the Chinese language. Yang et al. (2023) conducted a systematic review exploring such tools. In the 29 studies that they identified, 11 included data on language learners rather than L1 Speakers. The studies investigated corpora that ranged in size from 100 samples from a standardized L2 Chinese exam (the Hanyu Shuiping Kaoshi, HSK) to over 85,000 texts from L1 speakers of Chinese. The studies used a variety of metrics to evaluate the validity of scores produced from AES tools including Agreement Rate, Exact Agreement Rate, Pearson Correlation Coefficient, and Quadratic Weighted Kappa (QWK). The QWK scores ranged from 0.60 to 0.88, with the highest score for L2 learners reaching 0.714.

AES tools show great promise for L2 learners but to date they have been used with a very limited population for very specific purposes (e.g., mostly for English speakers on large scale, high stakes exams). As noted earlier, much of the research is dominated by large testing corporations who charge high prices for these assessments. Even when the costs are relatively low (~$4 per test) as is the case with ACTFL's new AES tool[1], testing groups of learners multiple times (i.e., the typical multiple assessments given in a language course or program) quickly increases the cost. This inevitably limits who can use AES tools and when and why they are applied. AES tools that are not developed and managed by large testing corporations often require high levels of technical and statistical expertise, which also limits who can use or develop these tools. In this study, we view the emergence of ChatGPT as a potential opportunity to explore a wider range of applications of AES for users with varying levels of technical and statistical expertise.

[1] https://www.actfl.org/news/actfl-and-lti-introduce-groundbreaking-automated-scoring-system-for-the-aappl-spanish-presentational-writing-component

2.2 ChatGPT and L2 writing assessment

Even in the first year since the release of ChatGPT, there have been many articles published on the applicability of using ChatGPT as an assessment tool. Most recently, Pfau et al. (2023) compared ChatGPT 3.5 Turbo's ability to identify errors with that of human raters using a corpus of essays at multiple proficiency levels produced by L1 Greek L2 English writers. They found that although ChatGPT did miss some errors, it was still strongly correlated with human raters (r=0.97). They note that even though human editing is still needed, ChatGPT greatly increases efficiency when identifying errors. Similarly, Jiang et al. (2023) also used ChatGPT in addition to three other AI tools to automatically identify errors in L2 Chinese writers. Similarly they found that AI models were very accurate with most of their models reaching around .8 accuracy levels. While being able to identify errors is important, it does not in itself lead to an assessment score.

In another study exploring the use of ChatGPT as an assessment tool for English language learners, Mizumoto and Eguchi (2023) used an IETLS TASK 2 rubric as the query (prompt) and used it to analyze 12,100 essays from the TOEFL11 test. The essays were previously rated by humans by separating them into either low, medium, or high levels on a five-point scale (following Blanchard et al., 2013), though little information is given on how these essays were scored. Mizumoto and Eguchi found that while ChatGPT had acceptable levels of reliability (quadratic weighted kappa~=0.38), a number of other statistical measures (e.g. GPT scores + Lexical measures + Syntactic complexity measures, + others) were needed to improve the scores to a QWK of .6. While this is promising, it again highlights the technical expertise needed to achieve accurate and reliable scores, thus undercutting a major affordance of tools like ChatGPT.

It is important to note that both studies only used and evaluated one prompt in their analyses and involved advanced English language learners (similar to other studies on AES tools). Further, there was no mention of the temperature parameters in either of these studies. These are not trivial points as they can impact the outcome of a query in ChatGPT significantly. Temperature in ChatGPT is a value between 0 and 1 that reflects the amount of variance or randomness that is allowed in a response to a prompt. The default setting is 0.7 which is argued to be the ideal setting for generating human-like text. This is somewhat problematic for assessments as scores given by ChatGPT will vary depending on the temperature level. For example, in Mizumoto and Eguchi's (2023) study, they noted that when running the same analysis twice their scores varied. Ultimately, they argued that this variance was acceptable, but if they had lowered or raised the temperature level, their reliability score between the two scores would undoubtedly follow suit. Thus, it is not unreasonable to assume that their results will vary at different temperature settings as it has in other studies (Coyne et al., 2023). While having a lower temperature may be ideal for returning numeric values, having a higher temperature may be needed when getting qualitative feedback or details on errors in a sentence as was the case in Pfau et al. (2023).

Coyne et al. (2023) also explored the use of ChatGPT as an assessment tool with English data that included errors. They were interested in exploring how well ChatGPT engaged in grammar correction. The authors identified 20 English sentences with errors and then explored how ten different prompts performed in identifying the grammar errors compared to human raters. They found that overall GPT-4 performed well in identifying errors and tended to perform better at lower temperatures. Equally important they illustrate that prompt engineering, the iterative process of developing effective prompts for generative AI, is an important factor in determining the effectiveness of ChatGPT as an assessment rating tool. With a temperature of .1, their prompts ranged in GLEU scores (a metric for error correction) from 0.31 to 0.582 across different prompts.

In this paper we argue that studies exploring ChatGPT should report both temperature and prompting strategies. But more importantly, we should explore the use of ChatGPT in a way that aligns with the affordances provided by the tool. Therefore, we argue that accuracy and reliability can be increased with effective prompting strategies. OpenAI has suggestions for improving prompting strategies, such as including more details in queries for relevant answers, asking ChatGPT to take on a role, using delimiters to indicate distinct parts of the prompt, specifying steps required to complete a task and asking the model to reflect on those, providing examples, and specifying desired output length (https://platform.openai.com/docs/guides/gpt-best-practices). In the next section we highlight the potential affordances of automated assessments and generative AI specifically as they do (and might) relate to L2 classrooms and discuss the practical implications of these tools for such contexts.

2.3 Evaluating ChatGPT for classroom-based assessments

A number of frameworks have been developed to assess and evaluate the use of AES tools (e.g. Williamson et al., 2012). These frameworks generally focus on constructing relevance and representation, accuracy of scores, generalization, extrapolation, and use of scores (e.g. Enright & Quinlan, 2010). Given that these areas of focus all depend on the use of score, and subsequently the consequences and impact of a score, it is reasonable to first explore this area and move backwards. Ferrara and Qunbar (2022) note when discussing validity claims for AES, we must explicitly delimit the scope of the claims to be made about an assessment. In other words, in order to make a claim about the appropriateness of the inferences derived from a particular assessment, one must first clarify the type and nature of the assessment.

In our study, we are specifically considering the use of ChatGPT for classroom-based assessments. Classroom-based assessments are, simply put, assessments that are conducted in a classroom setting by a teacher (as opposed to, for example, large-scale standardized assessments). Exploring the potential role of using automated assessments in the classroom setting requires that we first explore potential needs that such tools can fill. Classroom-based assessment includes weekly quizzes, unit tests, exit tickets, among others.

Although classroom-based assessments are usually described as either being formative or summative in nature, that is, *for* learning and *of* learning, respectively, Black and Wiliam (1998) argue that formative and summative are not properties of assessments *inherent* to the assessments themselves, but rather are properties of the *uses* of the information gathered from assessments. In other words, inferences, conclusions, and data can be *used* formatively or in a summative manner, even with the same assessment. Additional use of assessments in language learning programs include for diagnostic purposes and/or placement testing, but these usually occur outside of the classroom setting by a program coordinator or administrator.

Assessments that are used for formative purposes tend to involve more qualitative feedback rather than simply providing a learner with a score. This is because assessments that are used formatively aim to improve learning rather than simply measure it. This would suggest research involving the effectiveness of an automated assessment system that is targeting formative skills should focus on how well and relevant the feedback given by the system is. In contrast, summative assessments tend to have an accountability and/or administrative role in education. These assessments come at the end of an instructional unit or course and provide evidence of the extent to which learners have achieved established goals. Because information from summative assessments is often passed to other stakeholders (e.g. parents and administrators), quantitative evaluations are used for ease of communication and convenience. These assessments tend to involve higher stakes as scores usually impact learner grades and thus these assessments may have a gatekeeping effect (Winke, 2021). Thus, for automated scoring that is being applied to summative assessment data, the focus should be on reliability and accuracy of the tool's ability to generate a score.

In the present study, we are exploring ChatGPT's capacity to assess Chinese L2 writing samples reliably and accurately. We specifically consider how language educators may make use of this tool in their classroom settings and thus we explore approaches that are practical for in-class implementations. Given that we are focusing on primarily the accuracy of ChatGPT's ability to generate a proficiency score (a summative use of assessment), we are focusing on the potential use of this tool serve as a second rater or as a tool for learners to engage in self-assessment practices.

With this mind, we consider measurements for confirming accuracy and reliability of scores generated by automated assessments. Williamson et al. (2012) argued that for high stakes assessments at ETS, their threshold for accuracy using a quadratic weighted kappa measurement (QWK) was 0.7. Automated assessments at ETS include the GRE and TOEFL, among others. These are tests that usually cost individuals over $100 and have gatekeeping roles for graduate school (Winke, 2021). Compared to classroom-based assessments, these have significantly more impact on one's future and thus while 0.7 is a good benchmark for evaluation it is reasonable to consider a lower threshold for classroom-based assessments. It is also important to note that writing topic or task has also been

shown to impact writing outcomes (James, 2008), and thus it is important to confirm that scores are reliable across writing tasks.

Finally, when considering relevance and representation, one must consider how scores are derived and how they map onto constructs that are being measured. In traditional automated assessment models score generation are quite intuitive. AES tools usually have a set of text metrics generated by Natural Language Processing techniques that represent parts of the writing construct. For example, Quinlan et al. (2009) provide a detailed overview of how 30 different indices (e.g. fragments, run-ons, proper nouns, etc.) map onto 8 subconstructs (e.g. Grammar, Usage, Mechanics, Style, Organization, Development, Lexical Complexity, and Topic-specific vocabulary usage) and further how those subconstructs are connected to writing standards. This is somewhat problematic with ChatGPT and other LLMs given that there is less transparency regarding how results are generated as they employ 'black-box modeling approaches' (Bauer & Zapata-Rivera, 2020, p. 24). In other words, one may ask ChatGPT to apply a rubric to a text (Mizumoto & Eguchi, 2023) or to generate similar metrics as found in other AES studies (e.g. count the number of fragments), but it is unclear how such metrics are actually calculated or how a rubric is applied (or not) to a text. While we cannot directly address this issue in this study, it is important to acknowledge when investigating the reliability and accuracy of ChatGPT as an assessment tool.

Thus, our study is guided by the following research questions:

1. How do prompting strategies affect the accuracy of ChatGPT generated scores compared to human raters?
2. Are ChatGPT scores reliable across different tasks?

3. Methods

3.1 Data set

Data from the present study were taken from a corpus of third semester university L2 Chinese learners (n=48) from a private university in the United States. As part of their regular coursework, these students completed a standardized L2 proficiency assessment of listening, speaking, reading, and writing during the final week of their semester. Students in this study ranged from a writing score of 4 (N=18) to 7 (N=6) on individual tasks (possible scores ranged from 1-9), corresponding to Intermediate Low and Advanced Low on the ACTFL proficiency scale, respectively. See Table 1 for a complete breakdown of students' scores on individual writing tasks by level.

Table 1 Frequency of Writing Scores by Human Raters

Score	ACTFL Proficiency Level	Counts
4	Intermediate Low	18
5	Intermediate Mid	57
6	Intermediate High	63
7	Advanced Low	6
Total		144

*Note each of the 48 students were scored on 3 writing tasks.

Data from the present study consists of each student's three writing tasks responses in this standardized assessment (n=144). The standardized assessment uses a computer-adaptive system, meaning that the difficulty level of writing task was determined based on their reading scores (computer-scored multiple-choice questions). There was a total of 9 possible tasks[2], 3 of which each targeted low-intermediate, intermediate, and advanced, respectively. Task level (intermediate-advanced) was determined by reading scores; task order was randomly assigned. The number of students who took each task at varying times (e.g. Time 1, Time 2, & Time 3) can be seen in Table 2. All students completed the tasks in the assigned order. Each writing task was scored holistically by one or two professional human raters and assigned a numeric score from 1-9, corresponding to Novice Low through Advanced High (CEFR levels A1 to C1) on the ACTFL scale. The present study, therefore, used the writing tasks, the students' responses, and the official assessment scores (from raters) to evaluate the efficacy of automated scoring using ChatGPT.

Table 2 Number of students assigned to each writing task

Prompt	Targeted Level	Time			
		1	2	3	Total
Newspaper	Intermediate	15	12	11	38
Lost in forest	Intermediate	12	14	12	38
Appliance	Intermediate	11	12	15	38
New pet	Low-Intermediate	4	2	2	8
Letter of appreciation	Low-Intermediate	3	1	4	8
Live anywhere	Low-Intermediate	1	5	2	8
Time in history	Advanced	1	1	0	2
Positive in hardship	Advanced	1	0	1	2
City council	Advanced	0	1	1	2

[2] Because the standardized test is a commercial test with copyright restrictions, the precise prompts cannot be shared here.

3.2 Data analysis

To assess the reliability of the scores generated by ChatGPT in this study, we use four reliability measurements including exact and adjacent agreement percentages, Pearson's correlation, and quadratic weighted kappa (QWK). Exact agreement percentage reflects the amount of exact agreement between the human rater and ChatGPT scores. Adjacent agreement percentages refer to scores by ChatGPT that were within 1 point (below or above) human rater scores. QWK is commonly used to quantify the degree to which measurements resemble each other (Williamson et al., 2012). Unlike correlation coefficients, QWK accounts for both correlation and agreement between measurements. In other words, while correlations may pick up on trends in similar directions, QWK also illustrates how close two scores are to each other. QWK is therefore more appropriate for assessing reliability than Pearson's *r* when there is systematic variability between raters or measurements for the same subject (Vanbelle, 2016). Another option for measuring interrater reliability is Cohen's kappa; however, this is limited to categorical ratings. Since the scores used in this study are ordinal numeric response options, QWK is more appropriate reliability indexes than Cohen's kappa. We report multiple metrics to ensure accuracy as suggested by recent studies (e.g. Doewes et al., 2023).

Additionally, to investigate fairness of ChatGPT in scoring these essays, we also use a mixed-effects regression to explore ChatGPT's scores across multiple writing tasks. Mixed-effects models are ideal when data are nested. In our study, we have participants who are scored on three different writing prompts at three different times. Given the likely effect of individual and time of writing (e.g. first writing task vs second or third writing task), we added these variables as random effect intercepts to the model. Additionally, we control for differences in proficiency and time spent on task by adding these variables as fixed variables. No interactions were added to this model. We first created a null model with only proficiency and time spent on the assessment entered into the model, and then we added a categorical variable for the writing topic. To make this variable more interpretable, we use effect coding which means that instead of having a reference variable with which to compare the effect of writing task, individual tasks are instead compared to a grand mean. These findings will be reviewed in the results section.

3.3 Technical considerations for analyzing text with ChatGPT

There are a few technical considerations that must be considered with ChatGPT. First, because we are analyzing 144 texts, it is not practical to use the browser-based platform for the analysis. Most users of ChatGPT simply navigate to chat.openai.com to submit a prompt. If we were to analyze our essays through the browser, we would need to copy and paste both a prompt and a text 144 times and then manually add scores to a database to be analyzed later. This would be a cumbersome process for us (and for any educator who is interested in using ChatGPT for assessment purposes). Additionally, for

assessment purposes we want to adjust the temperature on ChatGPT. This cannot be done through the browser.

In contrast to using a browser to submit prompts, ChatGPT also be accessed by using the Application Programming Interface (API) through a programing language like Python. Google Sheets has an extension that also allows users to access ChatGPT through the API[3]. This extension allows a user to query ChatGPT from within a spreadsheet. Figures 1 illustrates how one can define a cell using a call to ChatGPT. In the image 'prompt' refers to the message that will be sent to ChatGPT, value is the text that is to be analyzed, temperature receives a value between 0 and 1, and model refers to the version of ChatGPT that one wants to use. For this study we used GPT-4 and set our temperature to 0.1 to reduce variability of responses. By using a Google Sheet, we can upload all data including the text to be analyzed into one sheet. This can greatly increase efficiency when it comes to applying ChatGPT to multiple texts.

Figure 1 GPT in Google Sheets

It is also important to note that using the ChatGPT API in this way is not free and requires that users register with a credit card. GPT-3.5 Turbo costs $0.0015 (USD) per token for input, and $0.002 per token for output. While GPT-4 is significantly more at $0.03 per token for input, and $0.06 per token for output. Because our output is only 1 number, we are mainly focused on the cost of the input, which takes into account the length of the text the students write as well as the length of our prompt. Understanding the exact conversion from words to tokens is complicated because tokens are not directly related to letters or words, but rather to chunks of text. It is estimated that approximately 1000 tokens is equivalent to 750 words in English and about 1.7 tokens is equivalent to 1 character in Chinese. However, it is important to emphasize that these are estimates. For this reason, it is not possible to give an exact cost for each prompt analyzed, but to be transparent, we can report that we spent $76.03 to analyze 144 Chinese texts 10 times (for 10 prompts) with an average of 305 Chinese characters per text analyzed. Our prompts ranged from 298 characters to 6367 characters (including both English letters and Chinese characters) with an average of 1007 characters. This cost comes to approximately $7.60 per prompt or about $0.05 to analyze one text. Notably, OpenAI recently changed the cost of API use and has

[3] https://workspace.google.com/marketplace/app/gpt_for_sheets_and_docs/677318054654

reduced costs by half. The prices we report here reflect the pricing structure at the time of analysis (September-October 2023).

3.4 Prompt engineering

Similar to Coyne et al. (2023), we used 10 prompts (see Appendix) to explore how unique ChatGPT queries result in different outcomes for each student's test responses. In our first prompt, we start by asking ChatGPT to analyze student writings using the ACTFL scale without providing descriptions of the scale itself. We clarify that we only wanted a numeric value, returned. In our second prompt we become more detailed and provide simple descriptions for each individual proficiency level. In prompt three, we change to the AVANT descriptors (the developer and administrator of standardized assessment from which our data were collected). AVANT rubrics are based largely on ACTFL scales and descriptions, but they do use slightly different terminology. In prompt four, we apply a set of discrete rules that AVANT shared via presentation about their scoring procedures. This prompt relies on ChatGPT's ability to apply logical rules to essay scoring. In prompt five, we add the entire rubric from AVANT similar to what Mizumoto and Eguchi (2023) did in their study. In prompt six, we apply a specific strategy from OpenAI which suggests providing ChatGPT with a step-by-step procedure. In Prompts seven and eight we provide specific examples of what an essay at each level should look like. Prompt seven received one example and Prompt eight received two examples. Prompt nine is the same as prompt eight, except we used Chinese to prompt ChatGPT rather than English. Finally, prompt ten provides generic examples (e.g. not specific to the task) of each writing level.

Table 3 List and Descriptions of Prompts

Prompting Number	Prompting Strategy	Brief Description
1	Simple: No descriptions	Analyze using known knowledge about ACTFL scale
2	Simple: Apply Logic (ACTFL)	Add a description of each level
3	Simple: Apply Logic using AVANT descriptors	Add details from Avant
4	Simple: Rule-based: Avant	Apply clear cut off points
5	Complex: Complete Rubric from Avant	Complete Rubric
6	Complex: Detailed Step-by-Step Procedure	Step-by-step
7	Provide Examples: 1 Example	One-shot prompting
8	Provide Examples: 2 Examples	Two-shot prompting
9	Provide Examples: Same as P8 but in Chinese	Chinese Two-shot Prompting
10	Provide Examples: Generic Examples	Generic Examples

4. Results

Table 4 illustrates the findings from the ten prompts that we applied in our study. When using different prompts we found that correlations between ChatGPT and human rated scores ranged from 0.23 to 0.58. However, given the nature of the proficiency scales (i.e., an ordinal, nine-point scale), using the QWK is more appropriate for evaluating the accuracy of these prompts. The QWK scores range from 0.17 to 0.57 depending on the prompt used, with the most accurate scores coming from our 8th prompt. It is also important to explore the adjacent agreement given that these scores are on a nine-point scale. In other words, if a learner scores a 4 on the human rated assessments but receives a 5 from ChatGPT, the difference is between an Intermediate Low and an Intermediate Mid, this is not terribly concerning given that most students are assumed to be operating at a level above or below their proficiency level due to a number of factors (see Clifford, 2016, for discussion). In terms of adjacent agreement, we found a range between 74.3% and 97.2% with Prompt 2, 6, 7, 8, 9, and 10 all scoring over 90%.

Table 4 Similarity Measures

	Prompts									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Exact Agreement %	27.1	47.9	10.4	7.6	33.3	50.7	52.7	49.3	41.7	47.9
Adjacent Agreement %	74.3	97.2	45.8	40.3	86.8	92.4	96.5	93.8	95.8	95.8
Pearson's Correlation	0.23	0.45	0.42	0.53	0.54	0.42	0.49	0.58	0.50	0.45
Quadratic Weighted Kappa	0.17	0.44	0.18	0.18	0.37	0.38	0.48	0.57	0.42	0.45

We also provide a visual (See Figure 2) of the correlation and QWK scores to illustrate an issue with using correlation scores to assess automated scored. The scores are ordered from highest QWK to lowest. Prompt 5 and 4 both have significant correlation scores over 0.5, yet their QWK scores are much lower than their correlation coefficients. This suggests that there is some consistency with how ChatGPT is applying scores, but that the scores are not aligning with the scales being used (e.g. ACTFL's 1-9 scale).

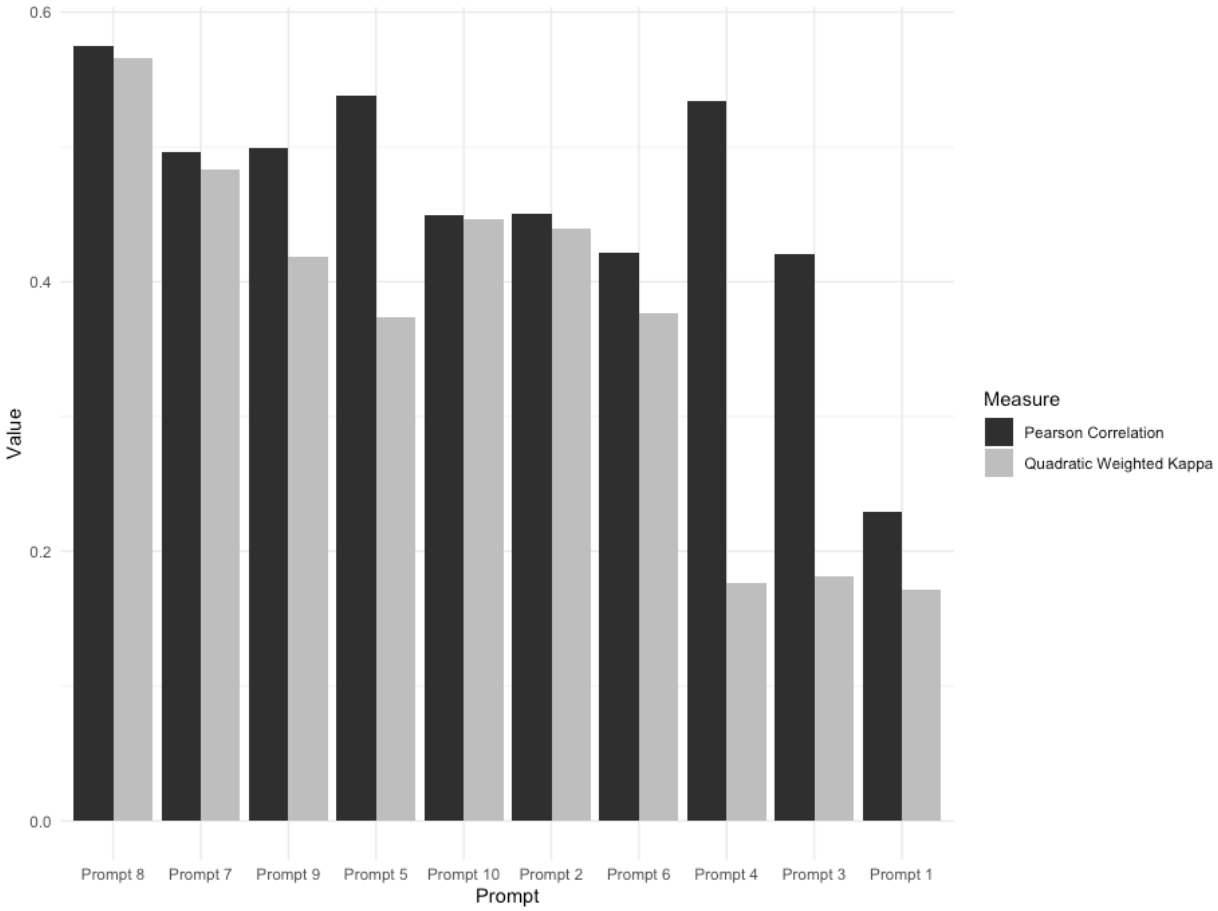
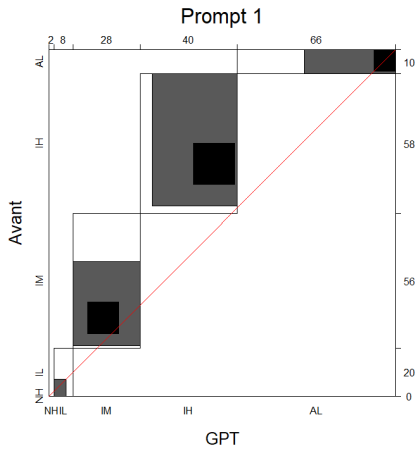
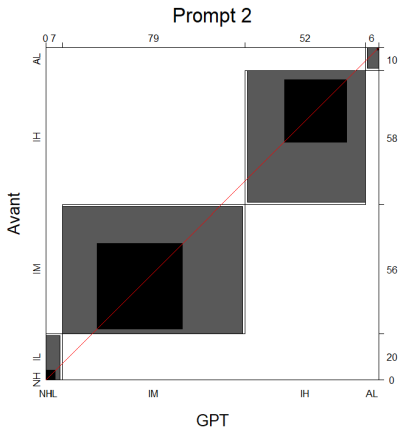
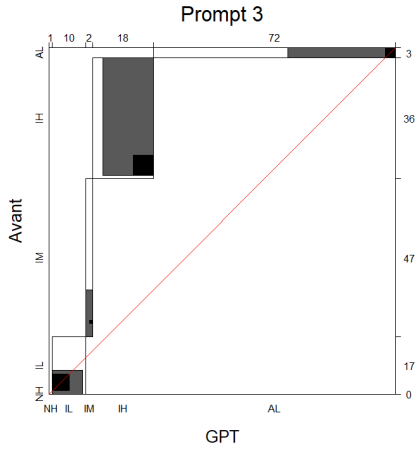


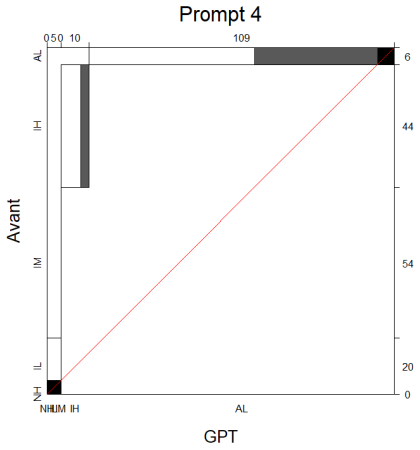
Figure 2 Comparison of Pearson Correlation and QWK for Each Prompt

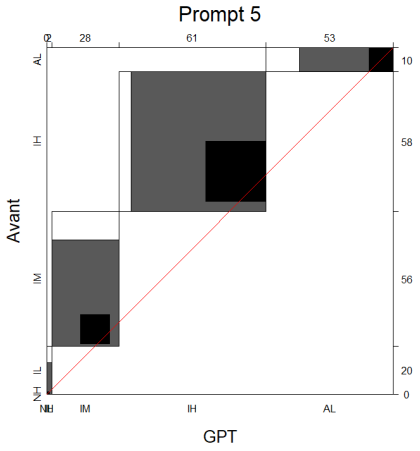
To continue exploring these prompts visually, we generated a series of adjacency plots for each prompt. For the visuals (Figure 3), black boxes represent an exact match between human-rated and ChatGPT scores. Grey boxes represent examples of a 1-point difference between human-rated and ChatGPT scores. White boxes represent examples that have a larger than 1 point difference between human-rated and ChatGPT scores. Thus, we are looking for visuals with large black boxes, smaller grey boxes, and even smaller white boxes. Furthermore, prompts that have boxes centered on the diagonal represent ChatGPT scores that are more closely correlated with human-rated scores.

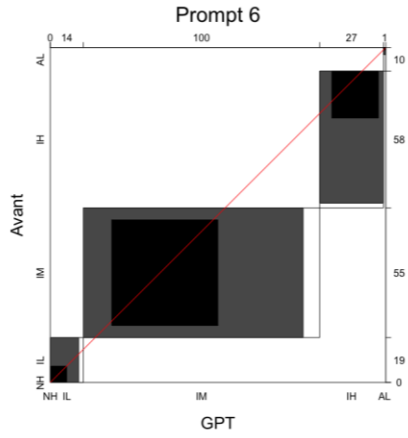

Prompt 1
Avant
GPT
NH IL IM IH AL


Prompt 2
Avant
GPT
NH IL IM IH AL


Prompt 3
Avant
GPT
NH IL IM IH AL


Prompt 4
Avant
GPT
NH IL IM IH AL


Prompt 5
Avant
GPT
NH IL IM IH AL


Prompt 6
Avant
GPT
NH IL IM IH AL

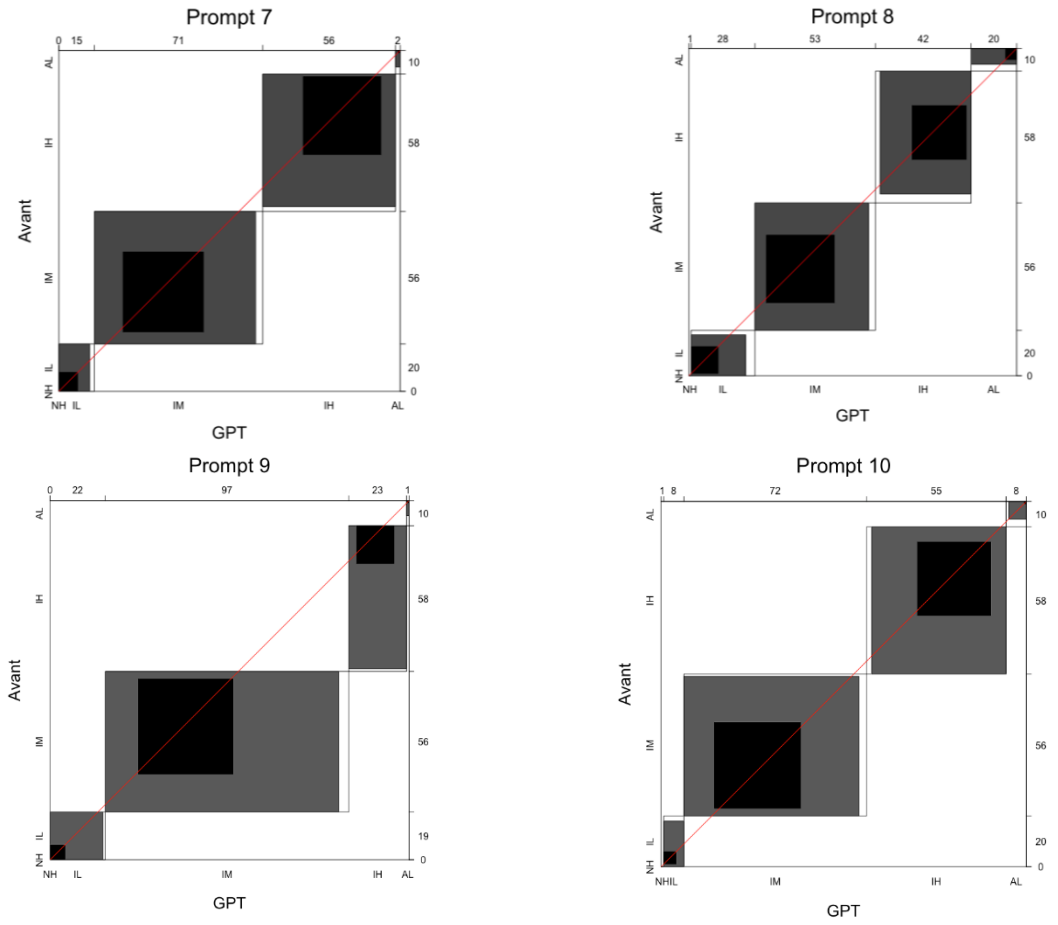


Figure 3 Exploring Prompt Performance via Adjacency Plots

Looking at Prompts 7, 8, it is clear that there are minimal examples of scores that diverge by more than two points, while prompts 3 and 4 are clearly problematic. Interestingly, Prompt 9, which is the same as Prompt 8 except it was written in Chinese performed worse.

To explore our second research question, we conducted a mixed-effects regression model to determine if ChatGPT scores are reliable across writing tasks. We used both individual participant and task order as random effects and compared the variance in random effects of individual and task between ChatGPT and human-rated scores. In both cases, individual differences account for large portions of the variance in scores with individual clusters accounting for ~26% of the variance in ChatGPT scores, and ~22% of the variance in human-rated scores. This is reasonable since we have varying proficiency levels in our data set. The variance associated with order of tasks is moderate in both cases at ~8% and ~6% respectively, but this does illustrate that task order plays a role in final scores. Further analysis shows that scores tend to decrease as order of task increases. This is likely due to a fatigue effect and further establishes the need for a mixed-effects model.

Table 5 Mixed-effects Regression Results

	ChatGPT (P8)		Human-Rated Score	
	Null	Full	Null	Full
Proficiency	0.677***	0.665***	0.772**	0.816***
	(0.146)	(0.195)	(0.111)	(0.151)
Time Spent on Assessment (minutes)	0.003	0.004	0.007	0.006
	(0.005)	(0.005)	(0.004)	(0.004)
Appliance		-0.558**		0.098
		(0.196)		(0.154)
Letter of Appreciation		-0.286		-0.030
		(0.325)		(0.258)
Live Anywhere		0.055		0.179
		(0.326)		(0.259)
Lost in Forest		-0.244		-0.002
		(0.196)		(0.154)
New Pet		-0.133		0.152
		(0.325)		(0.258)
Newspaper		0.261		-0.040
		(0.196)		(0.154)
Positive Hardship		0.658		0.742
		(0.511)		(0.413)
City Council		-0.258		-0.722
		(0.511)		(0.413)
Constant	1.457	1.630	0.764	0.510
	(0.857)	(1.105)	(0.650)	(0.853)
Observations	144	144	144	144
Log Likelihood	-187.748	-177.236	-148.214	-149.110
Akaike Inf. Crit.	387.497	382.472	308.427	326.220
Bayesian Inf. Crit.	405.316	424.050	326.246	367.797

Note: Topic is effect coded. *p **p ***p<0.001

Table 5 demonstrates that when controlling for proficiency and time spent on task, writing task does predict outcomes for ChatGPT while it does not for human raters. Interestingly, the 'appliance' prompt was associated with more than a half-point lower score compared to other prompts. This is not the case for the human rated assessments. These findings will be explored further in the discussion section.

5. Discussion

In this paper, we set out to explore the effectiveness of ChatGPT to automatically apply a rubric to Chinese L2 writers. To date we are unaware of other studies that have explored the use of ChatGPT to assess L2 Chinese writers other than Jiang et al. (2023) which primarily focused on error detection. More importantly, we position our research as addressing the potential practicality of using these tools in classroom settings. With this in mind, we considered the time, technical expertise needed, and cost of implementing AES tools. In terms of technical expertise and time, we acknowledge that any approach that requires developing expertise in statistical measures and/or software is unlikely to be integrated into mainstream teaching practices. Thus, we focused on unique prompting strategies that can impact the accuracy of ChatGPT to assess writings, which we argue that any teacher would be readily able to implement without extensive training. More specifically, we applied rubrics specifically designed for the writing samples to student writings automatically with the help of ChatGPT. In our prompting strategies, we kept the prompts short to reduce costs while also adhering to best practices provided by OpenAI. In our series of prompts, we were detailed yet concise, we added logical steps for ChatGPT to follow, we tried prompts in both English and Chinese, and we tried prompts that included examples of performance at each level of the rubric, all of which teachers could be readily expected to do for classroom-based summative assessments.

To answer our first research question, we discovered that prompting strategies have a profound impact on scoring accuracy. Our results show that Pearson's *r* correlation scores ranged between 0.24 and 0.57 and QWK scores similarly ranged between 0.17 and 0.58. These are large differences and were primarily due to how ChatGPT was prompted. If generative AI tools are to be used widely, it is clear that training users on how to prompt ChatGPT for assessment purposes is needed. Further, steps to ensure reliability and accuracy are also needed. In our study, we found that using multiple examples in lieu of detailed descriptions of levels in a rubric performed the best, however, even with our best prompt we noticed some discrepancies between ChatGPT and the human raters. When we explored the performance of the prompts more closely, we noted that some students' scores were more closely aligned with human raters, while others diverged more. To better visualize this we plotted each individual's writing scores on three different writing prompts (see Figure 4). Purple shading represents writing tasks in which both human-raters and ChatGPT scored participants exactly equivalently. For example, participant #142 was given a 5 on all three writing prompts by both human raters and ChatGPT. Examples like this are most ideal for making robust validity claims about ChatGPT as an assessment tool. The colors blue (human-raters) and pink (ChatGPT) indicate scores that were not overlapping. Thus, Participant 135, for example, was given a 5 on three of their writing samples by human raters and then two of these samples were given 4 by ChatGPT while the third score was given a 6. For participant 144, both ChatGPT and human raters scored one writing sample as a 7, while two samples were given a 6 by ChatGPT and two were given a 5 by human raters.

Figure 4 Rater and GPT Score Convergences and Divergences

We were not able to detect any trends between students who were scored poorly by ChatGPT in comparison to human raters. However with more data, identifying commonalities between students who were consistently scored incorrectly may be possible. Such data may provide insight into how ChatGPT is actually applying rubrics (i.e., help us dig further into its 'black box' mechanism).

Additionally, we explored the impact of writing topics on the reliability of our best prompt (prompt 8) visually. Figure 5 shows that most of the writing prompts came from either writing about an appliance that one finds to be useful (appliance), what one would do if they were lost in a forest (lost in forest) and one's perception about the relevance of newspapers in today's society (newspaper). The other topics had relatively fewer responses. Looking at the number of instances of exact agreement, ChatGPT seems to have performed better on the *newspaper* topic, while *appliance* and *lost in the forest* tended to have more

diverging scores. However, statistically only the *appliance* writing topic showed scores that differed significantly from other writing topics (being systematically lower by about half a point on the 9-point scale). Similar to our discussion above on how individuals were scored, exploring performance on prompts can also lead to valuable insight into how ChatGPT applies rubrics. For example, future research may want to extract all of the misclassified essays from the *appliance* prompt to determine if any themes emerge.

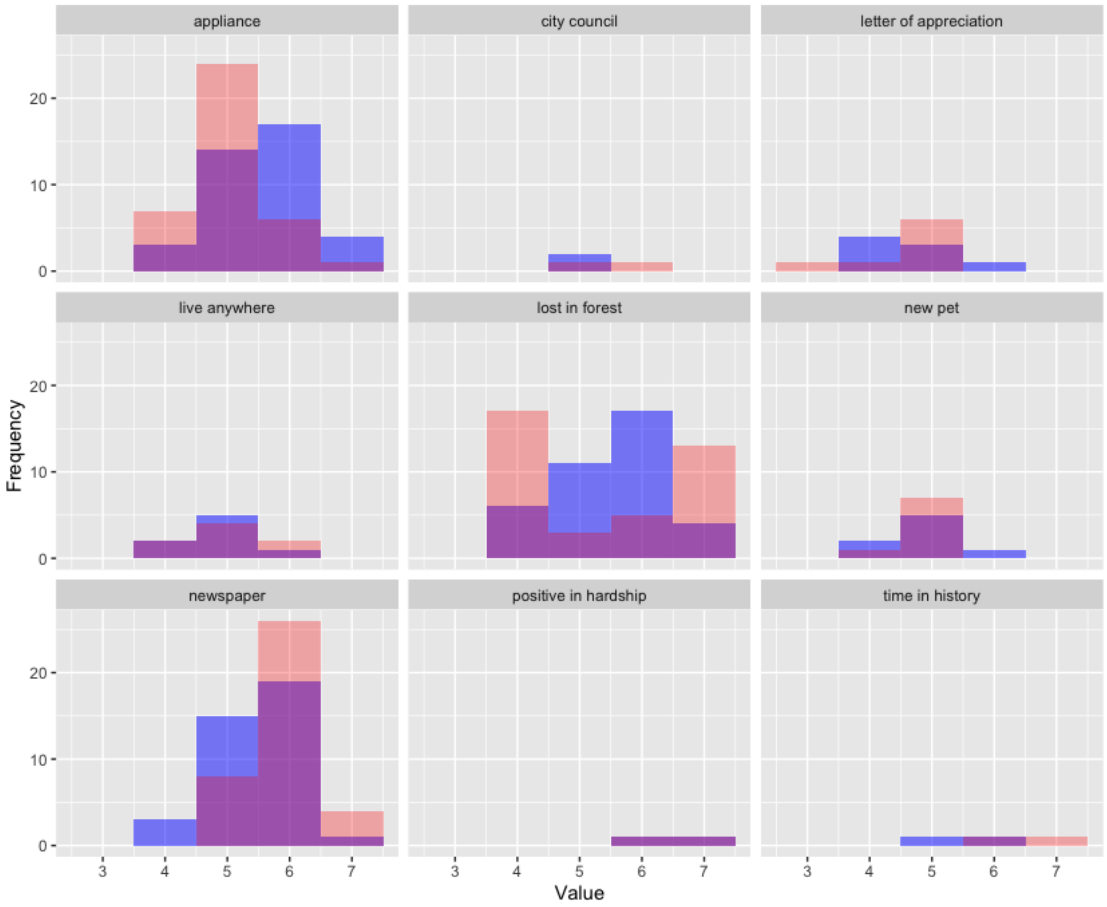


Figure 5 Rater and GPT Score Convergences and Divergences by topic

6. Conclusion

Our study did not find accuracy scores at levels reported in other AES studies. As noted earlier, many previous studies have generated scores that better approximate human-rated scores, with a number of studies finding QWK values over 0.7 (i.e., the threshold identified by ETS). That being said, many of the tools that have achieved or surpassed that threshold are either expensive or require technical expertise. Both of these caveats limit the widespread use of these tools. Additionally, it's important to note that many of the previous AES tools were created with a specific task and text type in mind. In our study, we applied one ChatGPT prompt to multiple writing tasks and found that scores were fairly reliable

across tasks. This is an important consideration for a classroom teacher who likely will not be able to customize their tool to each writing assignment.

Although our study did not find that ChatGPT reached a desirable reliability threshold, we still argue that it can be used as an assessment tool for certain cases in classroom-based assessments. The first and most obvious use case is as a second coder. ETS and other testing corporations often argue that AES tools should only be used as second coders (Ramineni & Williamson, 2018). Only a few testing companies rely primarily on an AES tool. Classroom teachers rarely have time to check scores or allow a second coder to check even a small portion of their graded papers (raising questions about reliability, especially for higher stakes classroom-based assessments like final course exams). Using ChatGPT as a second coder may help identify potential biases and/or errors for classroom-based assessments. As we noted in our study, many of our prompts were within 1-point of the human raters on a 9-point scale more than 90% of the time. As a second rater we argue that a 0.57 QWK with a +90% adjacent rater agreement is more than sufficient. For educators looking to use this tool, we suggest running an automated assessment with ChatGPT and then identifying any cases in which ChatGPT is more than 2 points off the human rater score. This does not automatically mean that the human rater was wrong, but it does provide a good starting point for reflecting on scores and further analyzing individual cases of highly divergent scores (including, possibly, prompting opportunities for further conceptual and analytical alignment within language programs or among colleagues).

In addition to using ChatGPT as a second coder, we also believe that it could be used as a self-assessment tool for language learners. Research has shown that writing in an L2 can benefit language learners (Polio & Park, 2016). However, teachers are often reluctant to assign writing without assessments. Using a self-assessment framework in which students write an essay, use ChatGPT to self-assess, and then reflect on the perceived accuracy of ChatGPT may not only increase the amount of writing that learners engage in but also it may support the development of metacognitive skills as well as digital literacy skills in relation to these new AI tools (Poole & Polio, 2024) as well as language proficiency literacy (see Coss and Van Gorp, forthcoming). Further, because this is used as a reflection tool rather than as a grading tool, any issue with accuracy is less concerning, as these can be mitigated by teacher-led or peer-to-peer discussion.

Regardless of how AI tools are used, our study highlights the importance of training teachers in how to best maximize both accuracy and reliability. The biggest takeaway that our study can offer at this point is that prompting matters. Luckily, there are easily-applied strategies that can greatly (relatively) enhance the reliability of ChatGPT-generated assessment score results. For example, the reliability scores in our study suggest that the best results come when a teacher uses past scored student examples or current examples to provide ChatGPT with an example of what writing looks like at each level. Prompts with

examples, therefore, may be the optimal strategy for maximizing the reliability of ChatGPT for the uses we have discussed here.

6.1 Limitations

There are a few key limitations to our study. First, we had a limited range of scores on the ACTFL scale and only 48 participants. Ideally, we would have had an equal number of participants at each level of the ACTFL scale with an equal distribution across writing tasks. That being said, our participants did range four levels on the ACTFL scale, and our sample is likely to reflect that of a foreign language classroom in which this tool may be used (i.e., Intermediate-level courses). Nevertheless, future studies should also explore the reliability and accuracy of this tool for novice and advanced learners. Secondly, we only explored 10 ChatGPT prompts, there are undoubtedly other ways of prompting this tool which may lead to better outcomes. Recently OpenAi has released updates that allow 'Plus' members to create their own ChatGPT that is customized to their needs. Creating a custom ChatGPT that has a database of learners past writings with human-rated scores may prove to be more accurate, reliable, and practical for language educators. Finally, we only explored one language, Chinese. It is likely that ChatGPT will perform better on these assessment tasks with languages that are better represented in ChatGPT's training data (e.g., English). To confirm this, future studies should explore variation in assessment accuracy across multiple languages. Finally, our study was focused on more summative uses of assessment evaluation. Future studies should examine the extent to which ChatGPT and similar tools are able to offer formative or diagnostic feedback, and the extent to which these tools could be incorporated systematically into language classrooms for these important, recurring purposes. In this line of research, the perceptions of stakeholder (students, teachers, etc.) would be important to explore concurrently with the accuracy and reliability of ChatGPT.

References

Attali, Y. (2015). Reliability-Based Feature Weighting for Automated Essay Scoring. *Applied Psychological Measurement, 39*(4), 303–313. https://doi.org/10.1177/0146621614561630

Bauer, M. I., & Zapata-Rivera, D. (2020). 'Cognitive Foundations of Automated Scoring', in Yan, D., Rupp, A. A., & Foltz, P. W. (Eds.). *Handbook of Automated Scoring*. CRC Press, pp. 13–28. https://doi.org/10.1201/9781351264808-2

Black, P., & Wiliam, D. (1998). *Inside the Black Box: Raising Standards Through Classroom Assessment*. GL Assessment.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2013). TOEFL11: A corpus of non-native English. *ETS Research Report Series* (i-15). https://doi.org/10.1002/j.2333-8504.2013.tb02331.x

Clifford, R. (2016). A rationale for criterion-referenced proficiency testing. *Foreign Language Annals, 49*(2), 224–234. https://doi.org/10.1111/flan.12190

Coss, M. D., & Van Gorp, K. M. (forthcoming). *What proficiency levels do K-16 world language learners achieve? An ACTFL Research Brief.* ACTFL.

Coyne, S., Sakaguchi, K., Galvan-Sosa, D., Zock, M., & Inui, K. (2023). Analyzing the performance of gpt-3.5 and gpt-4 in grammatical error correction. arXiv. https://doi.org/10.48550/arXiv.2303.14342

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing*, 28, 43–56. https://doi.org/10.1016/j.asw.2016.03.001

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7–24. https://doi.org/10.1016/j.asw.2012.10.002

Doewes, A., Kurdhi, N., & Saxena, A. (2023). Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. *16th International Conference on Educational Data Mining*. Germany.

Elder, C., Barkhuizen, G., Knoch, U., & Von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, *24*(1), 37–64. https://doi.org/10.1177/0265532207071513

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

Ferrara, S., & Qunbar, S. (2022). Validity Arguments for AI-Based Automated Scores: Essay Scoring as an Illustration. *Journal of Educational Measurement*, *59*(3), 288-313.

Huawei, S., & Aryadoust, V. (2023). A systematic review of automated writing evaluation systems. *Education and Information Technologies*, *28*(1), 771–795. https://doi.org/10.1007/s10639-022-11200-7

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, 5, e208. https://doi.org/10.7717/peerj-cs.208

James, C. L. (2008). Electronic scoring of essays: Does topic matter?. *Assessing Writing*, *13*(2), 80–92. https://doi.org/10.1016/j.asw.2008.05.001

Jiang, Z., Xu, Z., Pan, Z., He, J., & Xie, K. (2023). Exploring the role of artificial intelligence in facilitating assessment of writing performance in second language learning. *Languages*, *8*(4), 247.

Link, S., Mehrzad, M., & Rahimi, M. (2022). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*, *35*(4), 605–634. https://doi.org/10.1080/09588221.2020.1743323

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). https://doi.org/10.1016/j.resmal.2023.100050

Pfau, A., Polio, C., & Xu, Y. (2023). Exploring the potential of ChatGPT in assessing L2 writing accuracy for research purposes. *Research Methods in Applied Linguistics*, *2*(3). https://doi.org/10.1016/j.resmal.2023.100083

Polio, C., & Park, J. H. (2016). Language development in second language writing. In Manchón, R. M. & Matsuda, P. (Eds.). *Handbook of Second and Foreign*

Language Writing. de Gruyter, pp. 287–306. https://doi.org/10.1515/9781614511335-017

Poole, F. J., & Polio, C. (2024). From sci-fi to the classroom: Implications of AI in task-based writing. *TASK: Journal on Task-Based Language Teaching*, 3(2), 243-272.

Qian, L., Zhao, Y., & Cheng, Y. (2020). Evaluating China's Automated Essay Scoring System iWrite. *Journal of Educational Computing Research, 58*(4), 771–790. https://doi.org/10.1177/0735633119881472

Quinlan, T., Higgins, D., & Wolff, T. (2009). *Evaluating the Construct Coverage of the e-rater® Scoring Engine*. Research Report No. RR-09-01. Educational Testing Service.

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Ramineni, C., & Williamson, D. (2018). Understanding mean score differences between the e-rater® automated scoring engine and humans for demographically based groups in the GRE® general test. *ETS Research Report Series*, 2018(1), 1–31. https://doi.org/10.1002/ets2.12211

Reilly, E. D., Stafford, R. E., Williams, K. M., & Corliss, S. B. (2014). Evaluating the validity and applicability of automated essay scoring in two massive open online courses. *International Review of Research in Open and Distributed Learning*, *15*(5), 83-98. https://doi.org/10.19173/irrodl.v15i5.1857

Vanbelle, S. (2016). A new interpretation of the weighted kappa coefficients. *Psychometrika*, 81, 399–410. https://doi.org/10.1007/s11336-014-9439-4

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. 31st Conference on Neural Information Processing Systems. CA, USA: Long Beach, (Available at) https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning and Assessment*, *6*(2).

Wang, P. L. (2015). Effects of an automated writing evaluation program: Student experiences and perceptions. *Electronic Journal of Foreign Language Teaching*, *12*(1), 140–157.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational measurement: Issues and practice*, *31*(1), 2-13.

Winke, P. (2021). Foreword. In Mirhosseini, S. A. &De Costa, P. (Eds.). *The Sociopolitics of English Language Testing*. Bloomsbury, pp. vii–ix. https://doi.org/10.5040/9781350136025.0009

Yang, H., He, Y., Bu, X., Xu, H., & Guo, W. (2023). Automatic Essay Evaluation Technologies in Chinese Writing—A Systematic Literature Review. *Applied Sciences*, *13*(19). https://doi.org/10.3390/app131910737

