International Journal of TESOL Studies (2026)
2560213, 1-22 https://doi.org/10.58304/ijts.2560213

*Article*

# An Exploratory Evaluation of GPT-4's Consistency as an English Essay Rater: A Many-Facet Rasch Model Analysis of AI versus Human Rating Patterns

**Austin Pack***
**Steven Carter**
Brigham Young University-Hawaii, USA

**Alex Barrett**
Florida State University, USA

**Juan Escalante**
Brigham Young University-Hawaii, USA

**Mark Wolfersberger**
Brigham Young University, USA

Received: 1 January 2025/Received in revised form: 22 July 2025/Accepted: 30 December 2025/
/Available online: 20 March 2026

**Abstract**
This study examined the defensibility of using GPT-4 for automated essay scoring, using a Many-Facet Rasch Model analysis. Forty English for academic purposes student essays were rated by GPT-4 and four trained educators to assess nuances in rubric application, severity, leniency, and bias. Findings suggest that while GPT-4 tended to avoid the use of extreme scores, exhibiting a moderate central tendency rating, it does show a high level of consistency in its scoring behavior. This study contributes to understanding the extensions and limitations of using Generative AI tools in scoring essays, and provides insights into the use of AI tools in assessing writing.

**Keywords**
Generative AI, ChatGPT, artificial intelligence, automated essay scoring, assessment, education

## 1 Introduction

Writing is a fundamental skill in higher education that supports the development of language proficiency, critical thinking, and content learning (Behizadeh & Engelhard, 2011). While it is used in both formative and summative assessment, evaluating large volumes of student writing is labor-intensive for instructors

*Corresponding author. Email: Austin.Pack@byuh.edu

(Dikli & Bleyle, 2014). This challenge has driven interest in automated essay scoring (AES) for decades, dating back to early efforts such as Page's (1966) Project Essay Grade. Modern AES systems using machine-learning techniques have become reasonably reliable, enabling large testing programs to score hundreds of thousands of essays with consistency. Major exam providers, such as Educational Testing Service (ETS), routinely employ AES, and automated writing evaluation has even been integrated into commercial tools like Grammarly to provide instant feedback (Ding & Zou, 2024). Despite these advances, traditional AES solutions have limitations that hinder wider adoption in education. Testing companies do not publicly share their scoring algorithms, making those tools inaccessible outside specific exams, and commercial AES products are typically restricted to certain writing features and cannot be easily adapted to new prompts or rubrics (Ramesh & Sanampudi, 2021). In short, there remains a gap in accessible, flexible AES technology for general academic use.

Recent progress in generative artificial intelligence (GenAI) offers a potential solution to these AES accessibility and adaptability issues. Unlike conventional AES models that rely on predefined textual features and regression-based scoring, GenAI models like OpenAI's GPT-4 use large neural networks and transformer architecture to process text in a more holistic manner. These large language models (LLMs) are trained on massive corpora of text, enabling them to capture patterns of language and meaning beyond surface metrics (Stryker, n.d.). For example, GPT-based systems can consider an essay's content relevance, organization, and coherence in addition to grammar and vocabulary (Dai et al., 2023; Ding & Zou, 2024). Tools such as ChatGPT can be prompted with instructions or rubrics to evaluate a piece of writing and provide feedback. This suggests GenAI could overcome the "one-size-fits-all" nature of older AES. Instructors might repurpose a publicly available AI to score essays according to their own criteria, without needing a proprietary program. A critical caveat, however, is that models like GPT-4 were not originally designed explicitly for scoring student essays. As a result, there are open questions about their reliability and fairness. One key concern is whether an LLM's judgment might be biased or misaligned with either human expectations or any provided evaluation instruction (Pack & Maloney, 2024). This issue directly affects the validity of any decisions made based on AI-generated scores and warrants careful investigation before embracing GenAI as an assessment tool.

Using AI as an essay rater is still a relatively new practice, and researchers are only beginning to examine its implications. Traditional AES has become accepted in certain high-stakes contexts (e.g. college admissions and language proficiency exams), but relying on a black-box deep learning model for scoring has little precedent. Topuz et al. (2025) argue that deploying an AI as a rater may be acceptable with human-in-the-loop conditions. Nevertheless, even under such conditions, fundamental validity issues must be addressed. Chief among these issues is that raters apply the scoring criteria consistently and accurately for each task. In practice, this means well-trained human raters are expected to behave interchangeably; meaning each rater understands the rubric in the same way and produces scores in line with others. When raters are interchangeable and exhibit high intra- and inter-rater reliability, this facet of the assessment supports the overall validity of score-based interpretations (Eckes, 2015). If we introduce an AI into the rating process, the same expectation applies: the AI's scoring should be as reliable and criterion-aligned as that of a human rater for its scores to be defensible. This perspective reveals a need for research that is addressed by the present study. While GenAI models have demonstrated impressive language capabilities, it remains unclear whether a model like GPT-4 can function as a rater without compromising validity.

The purpose of this study, therefore, is to investigate the feasibility of implementation of Generative AI tools like ChatGPT in language centers and EAP programs that use in-house rubrics. More specifically, the study explores whether the assumptions requisite to a defensible validity argument can be satisfied when GPT-4 LLM serves as a rater in a specific English for Academic Purposes (EAP) writing assessment context. Central to these assumptions is how well GPT-4 can use a rubric to score English language learner essays. Essentially, GPT-4's scoring behavior should reasonably match (or improve upon) that of effective human raters in terms of consistency and adherence to the rubric. To explore this, we compare GPT-4's ratings to those of trained human educators on a common set of student essays. By analyzing agreement levels, reliability metrics, and rating patterns in depth, we aim to determine whether GPT-4 can act as an interchangeable rater alongside humans. This inquiry has important practical implications: if GPT-4 can perform comparably to human raters, it could significantly alleviate the burden of essay scoring for instructors, especially in resource-limited programs (e.g. Intensive English Programs) where conducting multiple independent ratings is often impractical. On the other hand, if notable discrepancies or biases are found in GPT-4's ratings, that would signal caution and the need for additional training or controls before GenAI is used for high-stakes writing assessment.

## 2 Literature Review

### 2.1 Background on Automated Essay Scoring

Traditionally, AES has relied on feature extraction to train scoring models that are based on large samples of student essays. Features that are extracted from these essays are quantifiable aspects of language that can be used as indicators of writing quality, such as grammatical accuracy, lexical variety, word count, and semantic similarity (Attali, 2013). The model is then trained to associate these features against a corpus of essays that are scored by humans in order to arrive at a final score. For example, research by testing agencies and independent evaluations report correlations in the acceptable range (often r ≈ 0.7–0.9) between automated scores and human rater scores on standardized essay tasks (Hussein et al., 2019; Ramineni et al., 2012; Shermis & Hamner, 2013). However, reliability in terms of correlation with human scorers is only tenuous (Perelman, 2014), and researchers argue that basing scores on quantifiable language features is limiting and ignores other aspects of writing quality such as argumentation or persuasiveness (Attali, 2013). Empirical work by Perelman (2014) underscored this issue, demonstrating that it was possible to fool certain AES programs with essays that were long and linguistically varied but essentially nonsensical. These essays received surprisingly high scores because they targeted the algorithms' preferred features. Strategic gaming of AES has been documented: adding excessive keywords or repetitive phrases can inflate scores from an algorithm that counts word occurrences, even though such tricks would be obvious to human readers (Lochbaum et al., 2013). In summary, feature-based AES offers efficiency and consistency, but it may suffer from construct underrepresentation and vulnerability to construct-irrelevant inputs. These limitations highlight why high automated-human score correlations, by themselves, do not fully establish the construct validity of AES.

Despite the limitations of early AES approaches, one clear advantage they offer is consistency. Machine scoring is not subject to fatigue, mood, or subjective bias in the way human scoring can be. By design, an algorithm will apply the same criteria to each essay every time. Human raters, in contrast, often struggle with consistency. Even experienced examiners show variability in their judgments due to both systematic biases and random "noise" (Kahneman et al., 2021). For example, one instructor might consistently be more lenient than another (a severity bias), or a rater might score an essay differently depending on subtle contextual factors like the quality of essays they read just before it (sequence effects). Research has documented that human raters interpret and weigh rubric criteria differently, leading to idiosyncratic scoring patterns (Eckes, 2008, 2015). These observations set the stage for why educators and researchers have long been intrigued by automated scoring, and by extension, why the advent of GenAI is seen as an opportunity to further improve on both consistency and scope of automated assessments.

### 2.2 GenAI in Essay Scoring: Potential and pitfalls

The rise of LLM-based AI systems like GPT-3.5 and GPT-4 has sparked new research into automated writing evaluation. GenAI models could potentially address some weaknesses of traditional AES, but they also introduce new uncertainties. On the positive side, LLMs bring an extensive knowledge of language that might allow them to assess higher-level writing features. Because they are trained on diverse real-world text, they can recognize elements of discourse structure, topical relevance, and even stylistic nuance that fixed-feature AES might miss. They can also be instructed to apply complex rubrics with multiple criteria. Early studies have indeed found that LLM-based scoring outputs often align reasonably well with human judgments (e.g., Mizumoto & Eguchi, 2023; Pack et al., 2024; Shin & Lee, 2024). Moreover, LLMs are widely accessible through cloud-based interfaces, meaning that educators and students can use these AI tools on their own data, customize the prompts, and iterate as needed, something not possible with sealed proprietary AES systems.

However, the notion of using GenAI for AES purposes may only partially address the weaknesses of existing scoring methods. One issue is that GenAI chatbots like ChatGPT are non-deterministic, using probabilistic generation, so the exact response can vary across runs even with identical input (Ouyang et al., 2023). In practical terms, this means if we ask ChatGPT to score the same essay on two different occasions, we might get slightly different scores or feedback each time. Indeed, Bui and Barrot (2025) observed this in their study: ChatGPT (GPT-3.5) produced inconsistent scores when the same essays were resubmitted for scoring on separate trials. Yamashita (2024) conversely noted that GPT-4's scores in his experiment were *extremely* consistent on repeated attempts, highlighting that model behavior may differ across versions and settings. Mizumoto and Teng (2025) highlighted that, when using GenAI as rating qualitative data, the best-performing model (DeepSeek-V3) achieved only moderate agreement ($\kappa = 0.68$), while other models (GPT 4o, GPT 3mini, Gemini, and Llama) demonstrated fair-to-moderate agreement ($\kappa = 0.37–0.61$).

A second issue is the "black box" problem (Bathaee, 2018). The internal workings of an LLM are highly complex and not transparent as to why a model produced a given output. With a linear regression AES model, one could at least identify which features contributed to a score; with GPT-4, we have no straightforward explanation for its scoring decisions on a particular essay. This opaqueness complicates validation and trust. For example, if an LLM consistently gives lower scores to essays on certain topics or by certain groups of students, it would be hard to detect and diagnose such bias without extensive analysis, since the model's reasoning is not easily inspectable. Furthermore, bias and fairness are serious concerns; an AI might inadvertently reflect biases present in its training data (Johnson & Zhang, 2024). In summary, GenAI offers enhanced linguistic capabilities and usability but poses challenges in ensuring reliable, fair, and interpretable scoring. These challenges form a major part of why research is needed: to empirically evaluate whether the benefits of LLM-based scoring can be realized without unacceptable trade-offs in validity.

English writing in its various forms (e.g. English as a second/foreign language, English for academic/specific purposes) may be particularly well positioned for AES. Established AES systems are highly proficient in assessing objective features of writing such as spelling, grammar, and punctuation but often struggle with assessing things like task completion and cohesion (Attali, 2013). The purpose of ESL writing assessment is typically to gauge language proficiency and not necessarily creativity or the quality of argumentation (Weigle, 2013). This is contrasted with writing assessments designed for expert English users that may not even mention grading criteria that examines language proficiency such as spelling or grammar. Furthermore, marking student writing for spelling and grammar is a tedious task for teachers that can easily be accomplished by a machine, saving valuable teacher time to concentrate on writing issues that require human insight. As mentioned earlier, many English testing services, such as the TOEFL, already employ AES due to the extremely high numbers of test takers who use these scores to qualify for university entry or workplace certification (Weigle, 2013). With LLMs being accessible for anyone with an internet connection, English teachers and students potentially stand to benefit greatly from using them for unlimited AES, with the understanding that these tools have both advantages and drawbacks (e.g., Xing & Saeed, 2025; Boonmoh & Kulavichian, 2025).

### 2.3 Recent studies on LLM-based scoring

Researchers have only recently begun to scrutinize how GenAI models perform as essay scorers, especially for L2 (second language) writing. Initial results are encouraging in terms of score agreement with human raters, though they reveal variability in consistency and potential biases. A number of studies in 2023–2024 have evaluated ChatGPT/GPT-4 on learner essays and found a reasonably strong correspondence between the AI's scores and human scores. For example, Mizumoto and Eguchi (2023) explored the use of OpenAI's GPT-3.5 model (text-davinci-003) to score 12,100 TOEFL essays written by test-takers of various L1 backgrounds. They prompted the model with IELTS Task 2 writing rubric descriptors and had it score each essay on a 0–9 scale. The AI's scores showed high agreement with the human ratings at a broad level: despite some 1–2 point discrepancies on individual essays, the model successfully distinguished low-, medium-, and high-proficiency essays in line with human evaluations. This suggests that the LLM captured the overall quality differences the rubric was meant to assess.

Shin and Lee (2024) developed a custom GPT-4-based chatbot and used it to score 50 English essays written by Korean high school English as a foreign language students. They provided the AI with a four-category analytic rubric (Task Completion, Content, Organization, Language Use) and sample responses for calibration. The results indicated high correlations between the AI's analytic scores and those of experienced human teachers across all rubric categories. In fact, the agreement was strong enough that by traditional standards, ChatGPT could be seen as a useful co-rater in that context.

Similarly, Pack et al.(2024) used GPT-4 to score 119 EAP student writing samples and found high levels of agreement between some LLMs (GPT-3.5, GPT-4, Claude 2, Bard) and human raters. These studies support the feasibility of using GenAI for AES purposes, particularly with rubric-based scoring. It is also noteworthy that researchers have started to enhance AI scoring through fine-tuning. Latif and Zhai (2024), for instance, fine-tuned ChatGPT on a large set of writing samples with known scores. Their fine-tuned model demonstrated improved accuracy in scoring new essays, yielding score correlations with human ratings on par with commercial AES systems.

Not all findings have been uniformly positive; some reveal limitations or inconsistencies of LLM-based scoring. In their study of 200 argumentative essays from an Asian EFL college corpus, Bui and Barrot (2025) found that ChatGPT's (GPT-3.5) holistic scores often did not closely match the scores given by an expert human rater. The Pearson correlation between the AI and human scores was notably lower than seen in other studies, and in many cases the AI's exact score deviated by more than one level. Moreover Bui and Barrot (2025) observed that the AI's scores were inconsistent across multiple scoring rounds. An essay might receive a 3 in one run and a 4 in another, undermining confidence in its use for any formal assessment. Another study by Geçkin et al. (2023) also reported relatively weak agreement: GPT-3.5's scores had only slight to fair correspondence with two out of five human raters on a set of L2 English essays, indicating the AI aligned well with some human judgments but not others. This variability suggests that LLM scoring may be sensitive to prompting methods or essay characteristics. Wetzler et al. (2024) raised a related concern about bias in AI scoring patterns. They found that ChatGPT 3.5 and 4 exhibited a form of proportional bias: it tended to be stricter than human graders on higher-quality papers and more lenient than humans on lower-quality papers. In practical terms, ChatGPT compressed the score range by pulling top essays down and pushing weak essays up (relative to human scores). Such a bias could have serious implications, such as stronger students being under-rewarded and weaker students over-rewarded if an instructor relied on the AI's grades. Wetzler et al.'s (2024) analysis suggests this might stem from the AI focusing heavily on language errors (penalizing advanced essays that still had minor mistakes) while giving the benefit of the doubt to poor essays that nonetheless used simple, correct language. Indeed, independent observations by Parker et al. (2023) support this notion: they noted that ChatGPT's grading of nursing students' essays was generally stricter than human grading on similar tasks, perhaps because the model zeroed in on surface-level correctness.

In sum, while many studies in a variety of contexts here have found high AI-human score alignment (e.g., Mizumoto & Eguchi, 2023; Pack et al., 2024; Shin & Lee, 2024), a few have highlighted inconsistencies and biases that urge caution. These mixed results underscore that simple correlation metrics may mask important differences in how an AI is applying a rubric. To fully ascertain the trustworthiness of LLM-based scoring, a more nuanced analysis of rating behavior is required, one that goes beyond aggregate agreement and examines when and how the AI's judgments diverge from human expectations. The present study takes up this challenge by performing an in-depth analysis of GPT-4's rating patterns using advanced measurement models.

### 2.4 Assessing rater performance with Many-Facet Rasch Modeling

More complex types of analyses are likely to yield greater insight into the patterns that GenAI exhibits when interpreting and applying criteria while rating writing. Such studies could serve as important contributions to arguments supporting the validity of decisions based on writing assessments in specific contexts when GenAI has been included in the pool of raters. Though studies have investigated interrater reliability, a more thorough and in-depth method for examining GenAI's rating patterns is the Many-Facet Rasch Model (MFRM).

Rasch models involve a unique approach to assessment wherein examinees' raw test scores are logistically transformed to scores on a linear, equal-interval scale (Eckes, 2019). The resulting scores are arguably more representative or accurate relative to one another than raw scores are. The application of Rasch techniques can be particularly useful in rater-mediated assessment scenarios—instances where human raters' judgmental and decision-making processes can have considerable impact on assessment outcomes (Eckes, 2019, p. 153). In these situations, the MFRM allows for controlling for raters' tendencies (severity or leniency, interpretations of the scale, etc.) when calculating examinees' Rasch-transformed test scores (Linacre, 1989). One might argue that factoring in raters' tendencies is unnecessary if interrater reliability is reasonably high. However, this assumption is problematic because high interrater reliability only guarantees that raters tend to rank individuals in similar rank orders. It does not guarantee that raters were similarly severe or lenient (Bond & Fox, 2015, p. 170), nor does it provide any nuanced information about raters' patterns. Some even argue that the whole notion of achieving acceptable interrater reliability is fraught (Eckes, 2015, p. 41). Ultimately, conducting MFRM on rater-mediated assessment data is far more insightful, and using output for grading is perhaps fairer.

Another meaningful advantage of analysis with MFRM is its potential for shedding light on the main effects of individual facets (i.e., raters, rubric categories, etc. (Eckes, 2015, 2019). For instance, it is possible to explore the unique tendencies that each rater may have and consider their effects on assessment outcomes. It is also possible to investigate interactions between facets, such as how individual raters interact with a scale or specific scale categories when compared to other raters. This is similar to differential item functioning (DIF) analysis but is called differential facet functioning (DFF) analysis (Eckes, 2019). When studying raters, ultimately, the purpose of these analyses is often to determine whether or not any rater or raters exhibit erratic or unique patterns of behavior.

With regard to evaluation and generalization of assessment, Knoch and Chapelle (2018) identified two specific warrants (and their associated assumptions) which relate directly to raters' consistency and accuracy: first, "raters rate reliably at [the] task level" (p. 483), and second, "different raters assign the same ratings to responses" (p. 488). Given the sudden and recent increased availability and accessibility of GenAI tools, studies that thoroughly explore their rating tendencies compared to trained raters are sorely needed to build evidence either in favor of or against these inferences in a validity argument. While previous studies have investigated AI as a rater, few known studies (Shin & Lee, 2024; Yamashita, 2024) have investigated AI's rating tendencies using MFRM analysis. Yamashita (2024) used GPT-4 to assign scores to 136 argumentative essays written by English language learners. Yamashita found

GPT-4's ratings showed a moderate alignment with learners' proficiency levels recorded in the corpus utilized, though they did not always correspond to human ratings. Shin and Lee (2024) leveraged GPT-4 to assess 50 English essays written by Korean secondary school EFL students and compared ratings with in-service English teachers. Results of MFRM analysis indicated that GPT-4 showed a slightly greater deviation from the model than its human counterparts. However, the non-deterministic and blackbox nature of GenAI means uncertainty persists and, given the skepticism surrounding ChatGPT and other GenAI tools in higher education (Sullivan et al., 2023; Fütterer et al., 2023), extensive evidence is needed before these tools can be confidently used for AES.

Therefore, to better understand the defensibility of using GPT-4's ratings for a specific learning context, this study examines differences in how GPT-4 and human raters score EAP student writing samples using a common rubric. We employ MFRM and other statistical measures to address the following research questions:

1. How do GPT-4's rating tendencies of EAP student writing compare to those of trained human language educators…

   a. in terms of severity/leniency and bias?

   b. in terms of overall nuance in scale application?

2. How does GPT-4's variation in interpretation of rubric criteria compare to the variation in interpretation by human language educators?

## 3 Methodology

### 3.1 Participants

Writing samples came from 40 matriculated university students (19 M, 21 F, 18-30 years old) enrolled in a general academic English writing course at a small liberal arts university in the Asia-Pacific region. Based on an institutional English placement test, students were assessed to be at a Common European Framework of Reference for Languages (CEFR) B1 (intermediate) level. These writings were rated by four native English-speaking, experienced EAP teachers, as well as a custom GPT that leveraged GPT-4 (version 11-06). These teachers had an average of 14.25 years of language teaching experience. Two hold PhD degrees in Applied Linguistics and English, and two hold MA degrees in Teaching English to Speakers of Other Languages.

### 3.2 Data collection procedures

Standard and institution specific ethical practices were followed in data collection and analysis procedures to ensure participant anonymity. Data came from EAP student responses to a writing prompt on a final exam. Participants were asked to write a fully-developed paragraph that demonstrates appropriate content, coherence in structure and ideas, appropriate language use (i.e., grammar, vocabulary, academic voice), and the integration of an academic reading source through quoting or paraphrasing.

All writing samples were rated by four teachers and GPT-4, using a rubric with four categories: content, coherence, language use, and sources and evidence (see Appendix A). Each writing sample received a score for each criterion, ranging from 1-10. To ensure inter-rater reliability in assessing student writing, these teachers calibrated six times throughout the academic year.

A custom GPT was created for the purposes of assessing writing with a specific rubric (Pack, 2023). Custom GPTs are "custom versions of ChatGPT that combine instructions, extra knowledge, and any combination of skills" (OpenAI, 2023). The custom GPT was given foundational knowledge documents, including the final exam reading passage that students were required to incorporate into their writing through quoting or paraphrasing, two example student writings with given scores for each criteria on the rubric, and the rubric used by teachers to assess student writing. In addition to this, the custom GPT was provided custom instructions, which are given in Appendix B. The instructions gave GPT-4 important contextual information, such as the students' proficiency levels and the writing prompts given to students, as well as objectives (use the rubric to evaluate the writing) and additional information that would normally be given to a teacher new to the course (such as how teachers interpret certain descriptors in the rubric, as agreed upon in calibration meetings). In using this custom GPT, student's writing samples were copied and pasted one at a time into ChatGPT's web browser interface, with a new chat being started for each student's writing.

### 3.3 Data analysis procedures

To conduct Many-Facet Rasch analyses of the data, the MINIFAC software (Linacre, 2024) was used. Three facets were specified: raters, examinees, and rubric categories. Our primary analysis utilized the Partial Credit Model (PCM) for the rater facet, allowing for a clear rendering of each rater's unique application of the rubric scale.

MFRM generates a large amount of data; therefore, various available model output and indices were leveraged to explore our research questions. To make the MFRM analysis more practical, our analyses were limited to the used portion of the 0 to 10-point rubric, converting it to a 0 to 6-point scale (i.e., scores of 4–10 on the rubric were rescaled to 0–6).

To answer RQ 1a and 1b,  the Wright map (Bond & Fox, 2015; Wilson, 2005) was used to explore each rater's severity/leniency, infit and outfit statistics, scale-category probability curves and their corresponding Rasch-Andrich thresholds (as dictated by their unique application of the rubric scale relative to examinees' final Rasch-generated scores), residual-expected rating correlations (rres,exp; Eckes, 2015; Myford & Wolfe, 2009; Wolfe, 2004; Wolfe & McVay, 2012), rater separation reliability and separation strata, and bias when rating individual examinees (raters*examinees).

To investigate RQ2, we conducted a secondary analysis specifying the PCM for both raters and rubric categories, thereby generating four unique applications of the rubric for each individual rater, one for each rubric category, enabling us to see how GPT-4 and each distinct human rater interacted with each rubric category when determining ratings.

## 4 Results

For a general sense of the ratings, the frequency of level occurrences by rubric category for each rater is given in Table 1.

### 4.1 The Wright Map and severity/leniency

The Wright Map (see Figure 1) is a vertical ruler which illustrates distribution of Rasch-transformed performance scores for examinees, distribution of severity and leniency scores for raters, and each rater's unique application of the rubric scale relative to examinees' scores all in one graphic representation. Examinees' raw scores, converted to scores on a linear, equal-interval scale are represented at the far left of the map.

Examinees are represented by asterisks. Those who are located vertically higher up in the Examinees column scored higher relative to lower examinees. On the other hand, raters who are located vertically higher up in the Raters column were generally more severe relative to the other (more lenient) raters.

According to the map, GPT-4 was the least severe rater. This finding is corroborated in Table 2, which shows the measurement results for the raters, with GPT-4's severity of -1.41 being lower than the severity of the other raters.

Table 1

*Frequency of Level Occurrences by Rubric Category by Rater*

| Rubric Category | Rubric Level | Rescaled Level | Frequency | | | | |
|---|---|---|---|---|---|---|---|
| | | | **GPT-4** | **RaterA** | **RaterB** | **RaterC** | **RaterD** |
| Content | 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 5 | 1 | 0 | 0 | 0 | 0 | 1 |
| | 6 | 2 | 0 | 0 | 0 | 0 | 4 |
| | 7 | 3 | 5 | 6 | 4 | 1 | 12 |
| | 8 | 4 | 14 | 10 | 11 | 20 | 6 |
| | 9 | 5 | 21 | 13 | 15 | 16 | 15 |
| | 10 | 6 | 0 | 11 | 5 | 3 | 2 |
| Coherence | 4 | 0 | 0 | 0 | 0 | 0 | 2 |
| | 5 | 1 | 0 | 0 | 0 | 0 | 0 |
| | 6 | 2 | 0 | 0 | 0 | 0 | 3 |
| | 7 | 3 | 5 | 2 | 3 | 3 | 4 |
| | 8 | 4 | 21 | 17 | 14 | 22 | 14 |
| | 9 | 5 | 14 | 14 | 9 | 13 | 13 |
| | 10 | 6 | 0 | 7 | 9 | 2 | 4 |
| Language Use | 4 | 0 | 0 | 0 | 0 | 0 | 3 |
| | 5 | 1 | 0 | 0 | 0 | 0 | 3 |
| | 6 | 2 | 0 | 0 | 0 | 1 | 2 |
| | 7 | 3 | 5 | 4 | 15 | 7 | 6 |
| | 8 | 4 | 32 | 28 | 16 | 23 | 10 |
| | 9 | 5 | 3 | 5 | 3 | 7 | 14 |
| | 10 | 6 | 0 | 3 | 1 | 2 | 2 |
| Sources & Evidence | 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 5 | 1 | 2 | 1 | 2 | 1 | 1 |
| | 6 | 2 | 1 | 0 | 0 | 1 | 4 |
| | 7 | 3 | 3 | 4 | 3 | 14 | 12 |
| | 8 | 4 | 32 | 16 | 13 | 20 | 11 |
| | 9 | 5 | 2 | 16 | 11 | 1 | 9 |
| | 10 | 6 | 0 | 3 | 6 | 3 | 3 |

Figure 1

*Wright Map of Primary MFRM*

```
----------------------------------------------------------------------------
|Measr|+Examinees|-Raters         |-Category | GPT |  A  |  B  |  C  |  D  |
----------------------------------------------------------------------------
+   3 +   High   +     Severe     + Difficult + (5) + (6) + (6) + (6) + (6) +
|     |          |                |          |     | --- |     |     |     |
|     |          |                |          |     |     |     |     |     |
|     |          |                |          | --- |     |     |     |     |
|     |          |                |          |     |     | --- |     |     |
|     |          |                |          |     |     |     |  5  |     |
|     |          |                |          |     |     |     |     |     |
|     |          |                |          |     |     |     |     |     |
|     |          |                |          |     |     |     |     |     |
|     |          |                |          |     |     |     |     |     |
+   2 + *        +                +          +     +  5  +     +     +     +
|     |          |                |          |     |     |     |     |  5  |
|     |          |                |          |     |     |     |     |     |
|     |          |                |          |     |     |  5  |     |     |
|     |          |                |          |     |     |     | --- |     |
|     | *        |                |          |     |     |     |     |     |
|     | *        |                |          |     |     |     |     |     |
|     |          |                |          |     |     |     |     |     |
|     |          |                |          |     | --- |     |     |     |
|     | ***      |                |          |     |     |     |     | --- |
+   1 + *        +                +          +  4  +     + --- +     +     +
|     |          |                |          |     |     |     |     |     |
|     | *        |                |          |     |     |     |     |     |
|     | **       |                |          |     |     |     |     |     |
|     | *        |                |          |     |     |     |     |     |
|     | **       |                |          |     |     |     |  4  |  4  |
|     |          |                | Lang Use |     |     |     |     |     |
|     | **       |                | Src & Ev |     |     |  4  |     |     |
|     |          |                |          |     |  4  |     |     |     |
|     | **       |                |          |     |     |     |     |     |
*   0 * *****    *                *          *     *     *     *     * --- *
|     | *        |                |          | --- |     |     |     |     |
|     | *        |                |          |     |     |     |     |     |
|     | ***      |                | Coherence|     |     |     |     |     |
|     | *        |                | Content  |     |     |     |     |  3  |
|     |          | RaterD         |          |     |     |     |     |     |
|     | **       |                |          |  3  |     | --- | --- |     |
|     | **       |                |          |     |     |     |     | --- |
|     | *        | RaterC  RaterB |          |     |     |     |     |     |
|     |          |                |          |     | --- |     |     |     |
+  -1 + *        +                +          + --- +     +     +     +  2  +
|     | *        | RaterA         |          |     |     |     |     |     |
|     | *        |                |          |     |     |     |     |     |
|     | **       |                |          |  2  |     |     |  3  | --- |
|     | *        | GPT-4          |          |     |     |     |     |     |
|     |          |                |          |     |     |     |     |     |
|     |          |                |          |     |     |     |     |     |
|     |          |                |          |     |     |  3  |     |  1  |
|     |          |                |          |     |     |     | --- |     |
|     |          |                |          |     |     |     |     |     |
+  -2 + *  Low   +    Lenient     +   Easy   + (1) + (1) + (1) + (1) + (0) +
----------------------------------------------------------------------------
|Measr| * = 1    |-Raters         |-Category | GPT |  A  |  B  |  C  |  D  |
----------------------------------------------------------------------------
```

Table 2

*Measurement Results for the Raters*

| Rater | Severity | *SE* | Fit Statistics | | | | Fair Average | Obs. Average | N of Ratings |
|---|---|---|---|---|---|---|---|---|---|
| | | | $MS_w$ | $t_w$ | $MS_u$ | $t_u$ | | | |
| GPT-4 | -1.41 | 0.13 | 0.77 | -1.5 | 0.74 | -2.0 | 4.12 | 4.10 | 160 |
| RaterA | -1.08 | 0.11 | 0.94 | -0.5 | 0.94 | -0.5 | 4.45 | 4.50 | 160 |
| RaterB | -0.78 | 0.11 | 1.01 | 0.0 | 0.99 | 0.0 | 4.34 | 4.40 | 140 |
| RaterC | -0.83 | 0.11 | 1.05 | 0.4 | 1.05 | 0.4 | 4.14 | 4.20 | 160 |
| RaterD | -0.51 | 0.08 | 1.19 | 1.5 | 1.21 | 1.7 | 4.04 | 3.90 | 160 |

*Note. $MS_w$ = mean-square infit statistic; tw = standardized infit statistic; MSu = mean-square outfit statistic; tu = standardized outfit statistic.*

## 4.2 Infit and Outfit Statistics, Probability Curves, Rasch-Andrich Thresholds, and Residual-expected Rating Correlation (*rres,exp*)

Infit and outfit statistics give insight into whether or not scoring tendencies for individual raters are erratic or follow expected patterns (e.g., higher scoring examinees generally receiving higher rubric ratings and vice versa from the rater in question). Fit statistics can also indicate how raters apply the rubric scale relative to one another: utilizing the full scale, exhibiting central tendency, etc. Standardized fit statistics (shown as *tw* and *tu* in Table 2) should ideally be close to 0. Positive values approaching or exceeding 2 indicate erratic behavior, whereas negative values approaching or subceeding -2 indicate overly predictable muted rating tendencies (e.g., central tendency; Bond & Fox, 2015). As is evident in Table 2, GPT-4 exhibits an over-fitting rating tendency with overly predictable, more central patterns of scoring. Further evidence of this can be seen in raters' probability curves (see Appendix C and D).

Probability curves help in illustrating at what points (relative to examinees' final scores) individual raters were most likely to apply the different levels of the rubric regardless of rubric category. For instance, examinees of ability levels 2.61 all the way to 6.00 (these are examinees' Rasch-transformed scores) were likely to receive a score of 5 (in rubric categories) from GPT-4, whereas examinees between ability levels -0.71 and 2.61 were likely to receive a score of 4 (in rubric categories) from GPT-4. These threshold values, called Rasch-Andrich thresholds (Linacre, 2024), are quite spread apart compared to the other raters' threshold values for score levels 4 and 5. It is worth mentioning that GPT-4's curves for points 1 and 5 are distinct, whereas points 2 and 3 appear to merge with 4.

What becomes clear on investigating all raters' probability curves is that the ability level at which most raters would award the maximum rescaled score of 6 (a score of 10 on the rubric), GPT-4 still awarded a 5 (9 on the rubric). In fact, the generative AI never awarded a single 6 (10 on the rubric) in any rubric category. It awarded a category score of 4 (8 on the rubric) the majority of the time (61.9%). The other levels it generally used were 5 (25.0%) and 3 (11.3%) (9 and 7 on the rubric). All other levels were used collectively less than 2.0% of the time. Though their tendencies differ slightly, the other raters' average use of the levels is more distributed. They used levels 3, 4, 5, and 6 respectively 16.2%, 40.4%, 28.0%, and 10.8% of the time (see Table 3 for more information).

The above data collectively indicate that, for this rating task, GPT-4 displayed some central tendency. The severity of this tendency can be further investigated by calculating the ***rres,exp*** for each rater. Central tendency causes raters to assign higher-than-usual scores to low-proficiency examinees, resulting in large and positive residuals. In contrast, high-proficiency examinees receive lower-than-usual scores, resulting in large and negative residuals. This rating pattern results in conspicuously negative residual-expected rating correlations (Eckes, 2015).

Table 3

*Frequency in Percentage of Rubric Level Use by Rater*

| Rubric Score | Rescaled Rubric Level | GPT-4 | RaterA | RaterB | RaterC | RaterD |
|---|---|---|---|---|---|---|
| 4 | 0 | 0.0% | 0.0% | 0.0% | 0.0% | 3.1% |
| 5 | 1 | 1.3% | 0.6% | 1.4% | 0.6% | 3.1% |
| 6 | 2 | 0.6% | 0.0% | 0.0% | 1.3% | 8.1% |
| 7 | 3 | 11.3% | 10.0% | 17.9% | 15.6% | 21.3% |
| 8 | 4 | 61.9% | 44.4% | 38.6% | 53.1% | 25.6% |
| 9 | 5 | 25.0% | 30.0% | 27.1% | 23.1% | 31.9% |
| 10 | 6 | 0.0% | 15.0% | 15.0% | 6.3% | 6.9% |

In the case of GPT-4, the ***rres,exp*** was .29. This correlation is clearly positive and the *t*-value (1.84) nearly surpasses the threshold for significance in the positive direction. Other raters' correlations fell between -.14 and .06. This provides some evidence in favor of GPT-4's rating pattern, contradicting other pieces of evidence that point toward central tendency, such as GPT-4's frequency in percentage of rubric level use (see Table 3 above), and GPT-4's rater probability curves (see Appendix C)

### 4.3 Additional insights into severity/leniency and bias

The general idea behind calibration and rater training is to encourage rating with negligible variation. In reality, however, rater differences are inevitable and training typically does not result in eliminating those differences (Eckes, 2015). Therefore, it is unsurprising that rater separation reliability for these data was .87 (the goal is the reverse—low reliability or relatively homogenous severity). Also, using the rater separation ratio (2.6) resulted in a rater separation index of 3.8, which means that raters could be separated into three "statistically distinct levels or classes of severity" (Eckes, 2015, p. 63). This spread in severity can be seen easily in Figure 1. Speaking of GPT-4 specifically, although it was clearly the most lenient, it was by no means an outlier in severity.

There were some instances of biased interaction when we investigated specific raters' evaluation of specific examinees. Numeric bias estimates (abnormally high or low ratings) can be used to calculate *t*-values and any absolute *t*-value exceeding 2 is considered indicative of substantive bias (Eckes, 2015). There were only nine such instances in these data, four of which involved Rater C. However, no systematic or correlated patterns of bias were evident.

### 4.4 Probability curves for distinct rubric categories by rater

The secondary analysis (specifying the PCM for both raters and rubric categories) yielded additional insight into how GPT-4 interacted with the differing categories on the rubric when compared to other raters. Appendix D presents raters' unique probability curves by each rubric category.

To reiterate, probability curves give a visual representation of the points (relative to examinees' final scores) at which individual raters were most likely to apply the different levels of the rubric. Of all the categories, raters awarded a larger range of levels and exhibited most variation on Sources and Evidence. GPT-4 followed this pattern to some degree, awarding some lower-level rubric scores (below 3) in this category. However, it is relatively clear that GPT-4's approach to rating had limited variation. In all other categories, it awarded only levels 3, 4, and 5 (7–9 on the rubric). As previously mentioned, GPT-4 refused to award a level 6 (10). This was true regardless of the rubric category. Essentially, across categories the only noticeable variation exhibited by the GenAI was in the span between the two thresholds which defined the upper and lower bounds of a level-4 (8) rating. While other raters followed expected patterns for human raters, GPT-4 was the only true anomaly.

## 5 Discussion and Implications

Given the rapid development and deployment of novel GenAI tools and general interest in how these tools might be used for language education and assessment purposes, the aim of this study was to better understand the defensibility of GenAI output for AES purposes—i.e., could evidence generally support the requisite warrants in a validity argument? We examined how OpenAI's GPT-4 LLM used a rubric to score writing samples of English language learners enrolled in a university level EAP writing course, and then compared its behavior to trained human raters.

We found that GPT-4 exhibited a predictable and mutable over-fitting central-tendency in rating behavior. Unlike human raters, it was unwilling to assign a score of 6 (10 on the rubric) for any category (content, coherence, language use, sources & evidence) and for any examinee. The finding that GPT-4 opted not to make use of the full range of scores on a rubric reinforces the findings of Pack et al. (2024) who found that LLMs, including Claude 2, Google Bard, GPT-3.5, and GPT-4, tended to refuse to make use of the higher and lower extremes of a holistic rubric's scale, instead opting for a central-tendency in rating. This trend was less extreme with human raters. In contrast, a study by Yavuz et al. (2024) found GPT-4 willing to utilize the full range of scores provided in their rubric for EFL essay grading, as did their human raters. However, they carefully chose essays that reflected a range of writing proficiencies which likely elicited the range in scoring.

In comparison with the four human raters, GPT-4 was the least severe. While Rater D was more erratic than other human raters, both human raters and GPT-4 generally gave scores in the 3–5 range (7–9 range on the rubric), which is to be expected given that these scores correspond to the rubric descriptors that most closely align with the expected intermediate proficiency level of the students. The only notable difference observed in GPT-4's performance across categories was the slight overuse and expanded probability-curve range of the rubric levels just below the maximum score (due to its refusal to give a maximum score).

These findings hold implications for the use of GenAI LLMs in language assessment. Wider adoption likely depends on a number of factors specific to learning contexts, including which LLM is being used (as their capabilities can differ drastically), how the LLM is fine-tuned or prompted, the type and complexity of the rubric being utilized, and the complexity of the writing assignment given to examinees. In instances where the LLM, rubric, and writing assignment differ from those used in the current study, we would strongly recommend educators and researchers conduct their own in-depth analysis of their chosen model's performance in rating before adopting and implementing it in actual assessment. Furthermore, there are ethical issues to consider (Pack & Maloney, 2023; Pack & Maloney, 2024), such as bias, maintaining student privacy, and adhering to government and institutional regulations (e.g., the Family Educational Rights and Privacy ACT or FERPA).

There is some evidence that supports an argument in favor of relying on GPT-4 as a rater of English language learners' writing. If educators do opt to use GPT-4 for assessing writing, it is important for them to understand the central-tendency currently exhibited by the model. One possible way to offset the central-tendency of GPT-4 could be to use a combination of GPT-4 and human rater generated scores to calculate a fair average by means of MFRM. Another, less costly, method could be implementing few-shot prompting techniques, where example writings which were awarded maximum and minimum scores in rubric categories by human raters are provided to the LLM as a reference.

### 5.1 Threats to validity and other limitations

The results of this exploratory study are vulnerable to some validity threats. First, as LLMs are stochastic, repeated scoring of each essay may have yielded variance in scoring behavior. The single pass technique used in this paper is reflective of how human raters might engage with essay scoring but can misrepresent LLM behavior. Future research should use a minimum of 5 to 10 replicates per essay.

Second, the custom GPT used in this study was seeded with the scoring rubric, reading passage, and example essays and scores. Although this improves face validity, it risks inducing central tendency by encouraging the LLM to reproduce the exemplar patterns rather than exercise independent discrimination. Future research should include ablation where zero-shot, rubric-only, rubric plus instructions, and the full custom GPT are examined in succession to isolate the effect of added context.

Another limitation in this study includes rescaling the rubric to increase the MFRM practicality. This may have contributed to the central tendency finding and therefore this result should be treated cautiously. Additionally, the sample in this study is typical of exploratory Rash analyses, but the single-exam rating from all B1-level students at one university means results should not be generalized.

## 6 Conclusion

This study extends the research on the use of GenAI for AES by exploring the feasibility of implementation in language centers and EAP programs using in-house rubrics. We also go beyond rater agreement and correlation for a deeper exploration of how a GenAI tool utilizes a rubric for AES. Despite the current constraints of GenAI tools, future models trained on data sets that are of greater quality and quantity will potentially perform better on language assessment tasks than extant models. As such, we see several avenues of research to further the field's understanding moving forward. First, as new models have been released (e.g., GPT-5), and will continue to be released, the rating behavior, and validity and reliability of these models' output for language assessment purposes needs to be continuously reassessed, especially given that many of these models will likely have improved reasoning skills. Also, existing models can be fine-tuned for language assessment purposes which may improve overall performance and potentially result in models abandoning their central-tendency rating behavior. Variations in prompting and parameter adjustments may also impact how LLMs utilize rubrics for AES. To conclude, GenAI technologies, such as LLMs, while nascent, hold promising potential for AES. Newer models and agentic systems will further the impact AI technologies have on language education and assessment.

## Ethics Approval

This study was approved by Brigham Young University-Hawaii's Institutional Review Board, approval number #23-55

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Appendix A – Writing Rubric

| Attribute | (0-5) Initial | (6-7) Emerging | (8) Developed | (9-10) Highly Developed |
|---|---|---|---|---|
| CONTENT<br>Writing displays effective communication through<br>• task completion,<br>• correct information,<br>• attention to context,<br>• analysis, and creativity. | Lacks a clear purpose and audience, and uses irrelevant, ineffective, incorrect, or confusing support. | Has a general sense of purpose with a vague audience, and employs some support that may occasionally be irrelevant, incorrect, or ineffective. | Has a fairly clear purpose and audience, and accomplishes the purpose with support that is mostly relevant. | Has a clear purpose and audience and accomplishes this purpose with effective and appropriate support. |
| COHERENCE<br>Writing displays appropriate organization related to<br>• formatting,<br>• paragraphing,<br>• purpose, and<br>• transitioning. | Organizes information in a manner that heavily interferes with the message.<br>Uses a layout and presentation that is very confusing or inappropriate. | Organizes information in a manner that requires occasional inference from the reader.<br>Uses a layout and presentation that is basic and sometimes confusing. | Organizes information in a manner that requires minimal inference from the reader.<br>Uses a layout and presentation that is clear and effective. | Flows from beginning to end using audience-friendly sequencing, transitions, and markers.<br>Uses a professional and appropriate layout and presentation. |
| LANGUAGE USE<br>Writing follows linguistics conventions such as<br>• spelling,<br>• punctuation,<br>• grammar, and<br>• word choice. | Uses language structures in repetitive, confusing, or inappropriate ways.<br>Frequently contains errors that interfere with meaning. | Uses language structures that are vague or general, and lack specificity or appropriateness.<br>May contain errors that interfere with meaning. | Uses some specific language structures but may have some problems with appropriateness.<br>Frequently contains minor errors that do not interfere with meaning. | Uses a wide range of specific and appropriate language structures.<br>May contain some minor errors that do not interfere with meaning. |
| SOURCES & EVIDENCE<br>Writing is appropriately supported by<br>• relevant reasons or examples,<br>• appropriate citations or source use, and<br>• disciplinary conventions. | Uses evidence in minimal, uneven, or confusing ways.<br>Attempts to use sources or referencing but these are ineffective or confusing. | Uses evidence that may be unclear, or that is relevant but inappropriate for the context.<br>Uses sources or referencing but these may be unclear, incomplete, or inappropriate. | Uses evidence that may contain superficial explanations or relevance.<br>Uses sources or referencing in a general manner but there be minor issues with consistency or professionalism. | Uses evidence in clear and effective ways.<br>Uses sources or referencing consistently and appropriately in a professional manner consistent with the context and discipline. |

## Appendix B – Instructions of the custom GPT "EIL317 Rubric Grader"

First read the EIL317 Writing Final Rubric, Final 1 Reading Passage, and example scored writings documents and make sure you understand them.

The user will input student writing. These students are ESL students at a B1/B2 CEFR level. After reading the rubric, assess the student's writing using the criteria in the rubric. First, give a score for content. Then give a score for coherence. Next give a score for language use. Lastly, give a score for sources & evidence. In each case, before giving a score, you should reflect on the criteria of the rubric. You do not need to justify your answers. Just provide a score for those four areas.

In making your judgements, consider what is expected of the students. Here are the prompts given to students:

"Choose ONE of the prompts below. Write 1-2 paragraphs in response to the prompt. You must write at least 300 words.  Make sure to include 1-2 quotes or paraphrases from the reading (with in-text citations).  Note: the title and author of the text are given in the instructions at the beginning of the reading.

What are the differences between Western and East Asian cultures? Which category do you fit in? Give evidence to support your answer.

How has your culture influenced your thinking and behavior? Make sure to use information and ideas from the reading to support your answer.

What are social norms in your own culture that you feel are good and should be embraced?  Which are not beneficial and should be excluded from your culture?  Make sure to write about 2 social norms in your own culture that you feel are good and 1 social norm that you feel is not beneficial.»

For content, a score of 8 or higher is on topic and developed. If the content is slightly off topic and the writing prompt is not fully answered, then a score of 7 should be given. Also, if the paragraph is not more than 250 words then it should not get a score of 8 or higher. If the writing is on topic and at least 300 words then it should get a score of 9 or 10. For coherence, a score of 8 or higher uses transition words to help the reader move from idea to detail, or idea to idea. For language use a score of 8 or higher is easily intelligible, a score of 9 or 10 uses academic words. For sources and evidence, students are expected to paraphrase or quote an article they are provided with. A score of 8 or higher means the student has quoted or paraphrased from the source correctly, a score of 9 or 10 means they have integrated it well into their writing. This article has been provided, it is called «Final Reading 1 Passage». If the student does not paraphrase or quote from the reading passage at all then they should not get a score higher than 5 for sources and evidence.

Again, do not justify your answers, just provide the scores only.

## Appendix C - Raters' Probability Curves

### GPT-4

```
   -6.0       -4.0       -2.0        0.0        2.0        4.0        6.0
    ++----------+----------+----------+----------+----------+----------++
  1 |111111111111111                                                    |
    |               1111                                           55555|
    |                   11                                      555     |
    |                     1                                  555        |
    |                      1                                5           |
  P |                       1                             55            |
  r |                        1           4444444         5              |
  o |                                  44       4       5               |
  b |                         1       4          44    5                |
  a |                                4             4  5                 |
  b |                          1    4               *5                  |
  i |                                              5 4                  |
  l |                           1  4              5   44                |
  i |                            3*33            5      4               |
  t |                           3*   33         5        4              |
  y |                          3  1    33      5          44            |
    |                         3 4        3   55             4           |
    |                       33 4   1      3*5                444        |
    |                      3  *2    1    55 333                 444     |
    |                 2*******  22222***5      33333               44444|
  0 |**********************5555555555 22********************************|
    ++----------+----------+----------+----------+----------+----------++
   -6.0       -4.0       -2.0        0.0        2.0        4.0        6.0
```

### RaterC

```
   -6.0       -4.0       -2.0        0.0        2.0        4.0        6.0
    ++----------+----------+----------+----------+----------+----------++
  1 |11111                                                              |
    |     111111                                                     666|
    |           11                                              666     |
    |             11                                            66      |
    |               1                                         66        |
  P |                1                                       6          |
  r |                 1                                     6           |
  o |                  1               4444                6            |
  b |                   1            44    44             6             |
  a |                               4        44          6              |
  b |                    1         4           4        6               |
  i |                        333334             45555556                |
  l |                     1 3    433           554    655               |
  i |                      *    4   3         5   4  6   55             |
  t |                     3    4     3       5     46      5            |
  y |                    3  1 4       33   55      64       55          |
    |                   *222 *          3 5       6  4        5         |
    |                22*    *2*         5*3     66    4        555      |
    |            222233   44   *2     55   33  6       44         555   |
    |      2222223333   44      1****5     66**33        4444        555|
  0 |*****************************6*************************************|
    ++----------+----------+----------+----------+----------+----------++
   -6.0       -4.0       -2.0        0.0        2.0        4.0        6.0
```

### RaterA

```
   -6.0       -4.0       -2.0        0.0        2.0        4.0        6.0
    ++----------+----------+----------+----------+----------+----------++
  1 |                                                                   |
    |1111                                                         666666|
    |    111                                                   666      |
    |       11                                               66         |
    |         11                                           66           |
  P |           1                                         6             |
  r |            1                                       6              |
  o |             1                                     6               |
  b |              11               4444444            6                |
  a |                1   33333     4       4          6                 |
  b |                 133     33  4         4        6                  |
  i |                 3         *4           4  555 6                   |
  l |               33 1       4 33           *5   *5                   |
  i |              3    1     4    3        55 4  6  55                 |
  t |             3      11  4      3      5    46     55               |
  y |           33         14        3   55     64       5              |
    |         33           41         335      6  4       55            |
    |       33           44  11       553     6    44       55          |
    |    333           44      11   55   33366       44       5555      |
    |3333          4444         5***    66663333       4444       555555|
  0 |***************************6666****11111111************************|
    ++----------+----------+----------+----------+----------+----------++
   -6.0       -4.0       -2.0        0.0        2.0        4.0        6.0
```

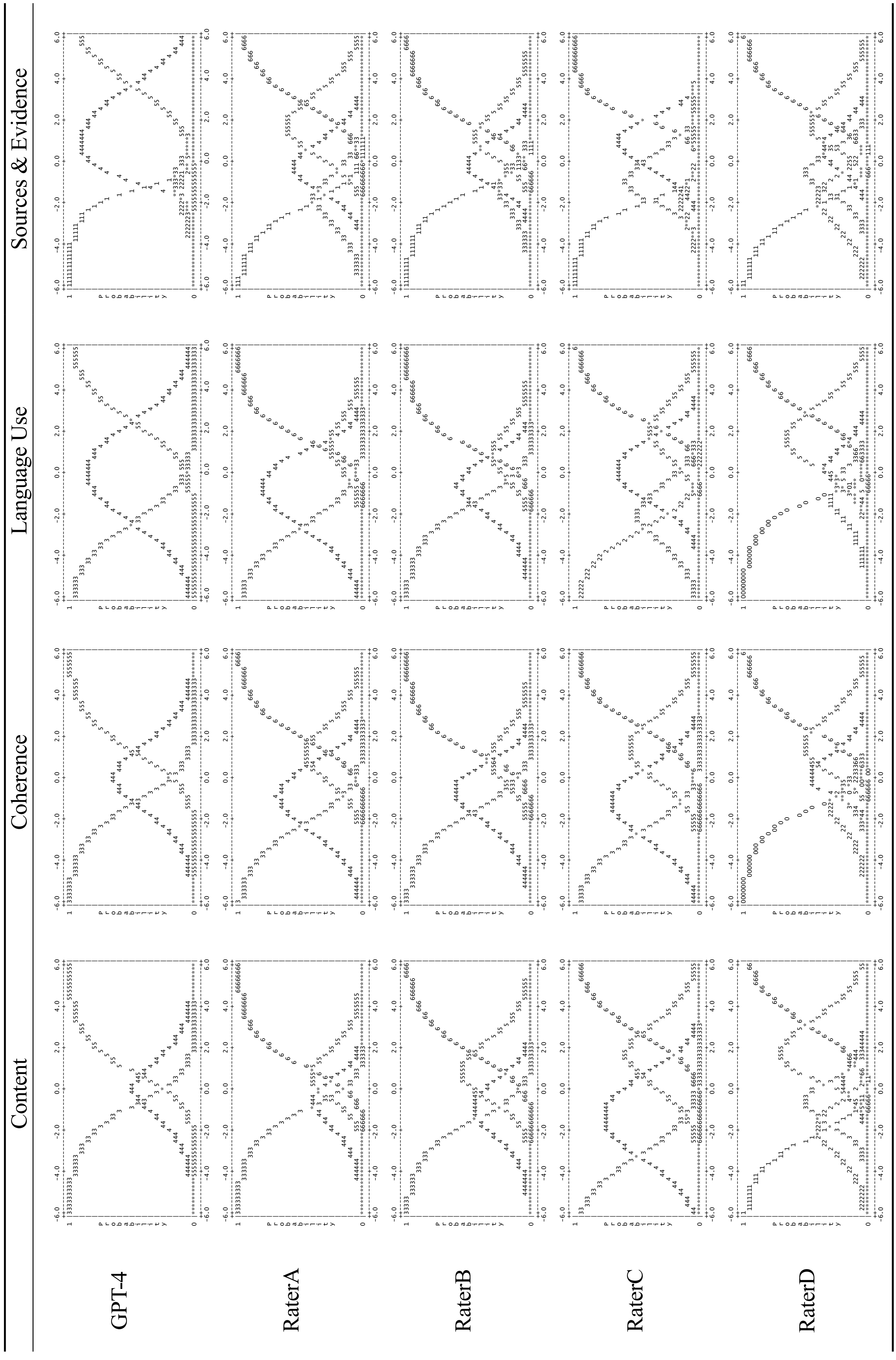
### RaterD

```
   -6.0       -4.0       -2.0        0.0        2.0        4.0        6.0
    ++----------+----------+----------+----------+----------+----------++
  1 |00000000                                                           |
    |        000000                                                 6666|
    |              000                                           666    |
    |                 0                                        66       |
    |                  00                                    66         |
  P |                    0                                  6           |
  r |                                                      6            |
  o |                     0                              66             |
  b |                      0                   5555     6               |
  a |                       0                55    55  6                |
  b |                                       5        56                 |
  i |                        0             5         655                |
  l |                              333    5         6   5               |
  i |                         0   3   344*4        6     5              |
  t |                          0 3   4435  44     6       55            |
  y |                         22*22 4  533   44  6          5           |
    |                     111*1*0  *2 5   3    46            55         |
    |                  111 22 3 1*4  *     3  6644             55       |
    |               111   2 33  44*15 22    **    444            555    |
    |        1111111  222233  44 55**1  2**6  333    4444           5555|
  0 |****************************6666***********************************|
    ++----------+----------+----------+----------+----------+----------++
   -6.0       -4.0       -2.0        0.0        2.0        4.0        6.0
```

### RaterB

```
   -6.0       -4.0       -2.0        0.0        2.0        4.0        6.0
    ++----------+----------+----------+----------+----------+----------++
  1 |                                                                 66|
    |1111                                                       666666  |
    |    1111                                                666        |
    |        11                                            66           |
    |          1                                         66             |
  P |           11                                      6               |
  r |             1                                    6                |
  o |              1                                  6                 |
  b |               1      33333                     6                  |
  a |                1   33     33                  6                   |
  b |                 133         3  444444        6                    |
  i |                 31          34      4       6                     |
  l |                3  1         443       45555*5                     |
  i |              33    1       4   3     554  6  55                   |
  t |             3       11    4     3   5   46     55                 |
  y |            3          1  4       3 5    64       5                |
    |          33            *4        5*3   6  44      55              |
    |        33            44 11      5   3 6     4       55            |
    |    3333            44     11  55    6*3       44       555        |
    |3333            4444       55**11 666   333     4444      5555555  |
  0 |***************************666666*111111111************************|
    ++----------+----------+----------+----------+----------+----------++
   -6.0       -4.0       -2.0        0.0        2.0        4.0        6.0
```

## Appendix D - Individual Raters' Probability Curves by Rubric Categories



## References

Attali, Y. (2013). Validity and reliability of automated essay scoring. In Shermis, M. & Burstein, J. (Eds.). *Handbook of Automated Essay Evaluation: Current applications and new directions* (pp. 181-198). Routledge.

Bathaee, Y. (2018). The artificial intelligence black box and the failure of intent and causation. *Harvard Journal of Law & Technology, 31(*2), 890-938. Retrieved from https://jolt.law.harvard.edu/assets/articlePDFs/v31/The-Artificial-Intelligence-Black-Box-and-the-Failure-of-Intent-and-Causation-Yavar-Bathaee.pdf

Behizadeh, N., & Engelhard, G., Jr. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing, 16*(3), 189–211. https://doi.org/10.1016/j.asw.2011.03.001

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Boonmoh, A., & Kulavichian, I. (2025). Enhancing Thai pre-service teachers' translation skills through AI tools, social media, and public feedback. *International Journal of TESOL Studies, 8*(2), 195-218. https://doi.org/10.58304/ijts.250906

Bui, N. M., & Barrot, J.S. (2025). ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. *Education and Information Technologies, 30*, 2041–2058. https://doi.org/10.1007/s10639-024-12891-w

Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y. S., Gasevic, D. & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pp. 323-325. https://doi.org/10.1109/ICALT58122.2023.00100

Dikli, S. & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing, 22*, 1-17. https://doi.org/10.1016/j.asw.2014.03.006

Ding, L. & Zou, D. (2024). Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence. *Education and Information Technologies*. https://doi.org/10.1007/s10639-023-12402-3

Ding, L., Zou, D., & Kohnke, L. (2025). ChatGPT as an automated writing evaluation tool: How students perceive it and how it affects their writing. *Education and Information Technologies*. https://doi.org/10.1007/s10639-025-13775-3

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155–185. https://doi.org/10.1177/0265532207086780

Eckes, T. (2015). *An introduction to Many-Facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang.

Eckes, T. (2019). Many-Facet Rasch measurement: Implications for rater-mediated assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative Data Analysis for Language Assessment Volume I: Fundamental Techniques* (pp. 153–175). Routledge.

Fütterer, T., Fischer, C., Alekseeva, A., Chen, X., Tate, T., Warschauer, M., & Gerjets, P. (2023). ChatGPT in education: Global reactions to AI innovations. *Research Square*. https://doi.org/10.21203/rs.3.rs-2840105/v1

Geçkin, V., Kızıltaş, E., & Çınar, Ç. (2023). Assessing second-language academic writing: AI vs. human raters. Journal of Educational Technology & Online Learning, 6(4), 1096-1108. http://doi.org/10.31681/jetol.1336599

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

Johnson, M. S., & Zhang, M. (2024). Using GPT-4o to score Persuade 2.0 independent items. *EdArXiv* preprint. https://doi.org/10.35542/osf.io/ctu8j

Kahneman, D., Sibony, O. & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little, Brown and Company.

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Validation of rating processes within an argument-based framework, 35*(4), 477–499. https://doi.org/10.1177/0265532217710049

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

Linacre, J. M. (1989). *Many-Facet Rasch measurement*. MESA Press.

Linacre, J. M. (2023). *A user's guide to FACETS Rasch-Model computer programs.* Winsteps. https://www.winsteps.com/a/Facets-Manual.pdf

Linacre, J. M. (2024). *MINIFAC computer program for many-facet Rasch measurement* (Version 4.1.6) [Computer software]. Winsteps. https://www.winsteps.com/minifac.htm

Lochbaum, K. E., Rosenstein, M., Foltz, P. W., & Derr. M. A. (2013). Detection of gaming in automated scoring of essays with the IEA. Paper presented at the *National Council on Measurement in Education Conference (NCME).*

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

Mizumoto, A., & Teng, M. F. (2025). Large language models fall short in classifying learners' open-ended responses. *Research Methods in Applied Linguistics, 4*(2), 100210. https://doi.org/10.1016/j.rmal.2025.100210

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement, 46*(4), 371–389. https://doi.org/10.1111/j.1745-3984.2009.00088.x

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

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. [Preprint from ArXiv]. https://doi.org/10.48550/arXiv.2308.02828

Pack, A. (2023). EIL317 Rubric Grader. https://chat.openai.com/g/g-nxTwzZxdM-eil317-rubric-grader

Pack, A., Barret, 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.* https://doi.org/10.1016/j.caeai.2024.100234

Pack, A., & Maloney, J. (2023). Using generative artificial intelligence for language education research: Insights from using OpenAI's ChatGPT. *TESOL Quarterly 57*(4), 1571-1582. https://doi.org/10.1002/tesq.3253

Pack, A., & Maloney, J. (2024). Using artificial intelligence in TESOL: Some ethical and pedagogical considerations. *TESOL Quarterly, 58*(2), 1007-1018. https://doi.org/10.1002/tesq.3320Page, E. B. (1966). The imminence of grading essays by computer. *The Phi Delta Kappan, 47*(5), 238-243. https://www.jstor.org/stable/20371545

Page, E. B. (1996). The imminence of grading essays by computer. *Phi Delta Kappan, 47*(5), 238-243.

Parker, J. L., Becker, K. & Carroca, C. (2023). ChatGPT for automated writing evaluation in scholarly writing instruction. *Journal of Nursing Education, 62*(12), 721- 727. https://doi.org/10.3928/01484834-20231006-02

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

Ramesh, D. & Sanampudi, S. K. (2021). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review, 55*, 2495-2527. https://doi.org/10.1007/s10462-021-10068-2

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T. & Bridgeman, B. (2012). Evaluation of e-rater® for the GRE® issue and argument prompts (ETS RR-12–02). *Educational Testing Service,* 2012(1). http://dx.doi.org/10.1002/j.2333-8504.2012.tb02284.x

Shermis, M. D. & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In Shermis, M. D. & Burstein, J. (Eds.). *Handbook of Automated Essay Evaluation*, (pp. 313-346). Routledge.

Shin, D., & Lee, J. H. (2024). Exploratory study on the potential of ChatGPT as a rater of second language writing. *Education and information Technologies, 29*, 24735-24757. https://doi.org/10.1007/s10639-024-12817-6

Stryker, C. (n.d.). What are large language models (LLMs)?. IBM. https://www.ibm.com/think/topics/large-language-models

Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning & Teaching.* https://doi.org/10.37074/jalt.2023.6.1.17

Topuz, A. C., Yıldız, M., Taşlıbeyaz, E., Polat, H., & Kurşun, E. (2025). Is generative AI ready to replace human raters in scoring EFL writing? Comparison of human and automated essay evaluation. *Educational Technology & Society, 28*(3), 36-50. https://doi.org/10.30191/ETS.202507_28(3).

Weigle, S. C. (2013). English as a second language writing and automated essay evaluation. In Shermis, M. D. & Burstein, J. (Eds.). *Handbook of Automated Essay Evaluation,* (pp. 36-54). Routledge.

Wetzler, E. L., Cassidy, K. S., Jones, M. J., Frazier, C. R., Korbut, N. A., Sims, C. M., Bowen, S. S., & Wood, M. (2024). Grading the graders: Comparing generative AI and human assessment in essay evaluation. *Teaching of Psychology, 53*(3), 298-304. https://doi.org/10.1177/00986283241282696

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46*(1), 35–51.

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice, 31*(3), 31–37. https://doi.org/10.1111/j.1745-3992.2012.00241.x

Xing, C., & Saeed, M. A. (2025). A systematic review on artificial intelligence (AI) technologies in ESL/EFL speaking skills. *International Journal of TESOL Studies, 8*(2), 240-270. https://doi.org/10.58304/ijts.250908

Yamashita, T. (2024). An application of many-facet Rasch measurement to evaluate automated essay scoring: A case of ChatGPT-4.0. *Research Methods in Applied Linguistics, 3*(3). https://doi.org/10.1016/j.rmal.2024.100133

Yavuz, F., Çelik, Ö., & Çelik, G. Y. (2024). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. British Journal of Educational Technology. https://doi.org/10.1111/bjet.13494

***Austin Pack*** is Assistant Professor of English Language Teaching and Learning at Brigham Young University-Hawaii. His research interests include the psychology of language learning and computer assisted language learning.

***Steven J. Carter***, MA TESOL, is an Associate Professor in the Faculty of Education and Social Work at Brigham Young University–Hawaii. He has taught English language courses in both intensive English and community English programs. He enjoys interacting with students from many cultures throughout the world and experiencing aspects of their cultures through his associations with them. His research interests include assessment and second-language reading.

***Alex Barrett***, PhD, is an instructional designer and researcher at Florida State University. His research centers on learner behavior in technology-supported environments, focusing on how learning and instruction are enabled and optimized. Specifically, he leverages learning analytics to investigate how human factors in computing— cognition, perception, memory, action— are influenced by the unique affordances of emerging technologies.

***Juan Escalante*** is an Associate Professor of English Language Teaching and Learning at Brigham Young University-Hawaii. His research interests include technology enhanced-language education, teacher training, and language assessment.

***Mark Wolfersberger*** earned a BA in Japanese Teaching and an MA in TESOL from Brigham Young University. After teaching for several years at an intensive English program, he studied under Rod Ellis at the University of Auckland and completed a PhD in Second Language Teaching and Learning. Since then, Mark has been working at Brigham Young University–Hawaii and Brigham Young University teaching academic English and training teachers. Although Mark is interested in many aspects of English language teaching, his primary area of professional interest is second language writing.