Research Methods in Applied Linguistics 4 (2025) 100199

Contents lists available at ScienceDirect

Research Methods in Applied Linguistics

journal homepage: www.elsevier.com/locate/rmal

Assigning CEFR-J levels to English learners' writing: An approach using lexical metrics and generative AI

Satoru Uchida [a,*], Masashi Negishi [b]

[a] Kyushu University, 744 Motooka, Nishi-ku, Fukuoka, Japan
[b] Tokyo University of Foreign Studies, 3-11-1 Asahi-cho, Fuchu-shi, Tokyo, Japan

ARTICLE INFO

Keywords:
Automatic essay scoring
CEFR-J
Generative AI
ChatGPT
CWLA

ABSTRACT

This study presents the CEFR-based Writing Level Analyzer (CWLA), a novel automated system for assessing English learners' writing proficiency. It assesses proficiency according to CEFR-J levels, a finely graded adaptation of the CEFR framework tailored to the English as a Foreign Language context, particularly in Japan. CWLA generates scores by combining vocabulary scores with AI-based analytical scores, allowing for a sophisticated, regression-based alignment with CEFR-J proficiency levels. By leveraging both straightforward lexical metrics and advanced AI scoring, CWLA provides accurate and interpretable assessments accessible via a user-friendly web interface. To evaluate CWLA's effectiveness, we conducted validation using the ICNALE GRA dataset, which showed a strong correlation of 0.88 between CWLA scores and human ratings. Additionally, entropy analysis indicated that CWLA's scoring distribution closely resembles human rater patterns, capturing the variability expected in human assessments. Three CEFR/CEFR-J specialists also performed expert validation, resulting in an agreement rate of 83.33 %, thereby supporting the system's alignment with expert judgment. These results suggest that CWLA is a reliable system for detailed CEFR-J level assessment, offering promising applications in language education and learner assessment across proficiency levels.

Introduction

Automated Essay Scoring (AES) is among the most sought-after technologies in the field of English language education. This demand arises from the importance of writing activities in English classes and the considerable burden of scoring them accurately, which often falls on instructors. At the same time, as Lumley (2002) points out, the process by which human raters assess writing is complex and not fully understood. Factors such as initial impressions and subjective biases can heavily influence evaluations; this indicates that writing assessments inherently contain subjective elements. AES, with its potential for high scoring accuracy, clear scoring criteria, and greater objectivity than human raters, greatly benefits English education. AES has the potential to provide immediate feedback, significantly reducing the burden on instructors.

Extensive research has been conducted on AES, and numerous systematic reviews are available to date (Ding & Zou, 2024; Hussein et al., 2019; Lim et al., 2021; Ramesh & Sanampudi, 2022; Uto, 2021). Furthermore, recent advancements in Large Language Models (LLMs) are reshaping the landscape of AES research (Mizumoto & Eguchi, 2023; Mizumoto et al., 2024; Yamashita, 2024; Yancey et al., 2023). In this evolving context, two major gaps persist: the alignment of AES results with CEFR-J levels, which provide a more granular

* Corresponding author at: Satoru UCHIDA, Kyushu University: 744 Motooka, Nishi-ku, Fukuoka, Japan.
E-mail addresses: uchida@flc.kyushu-u.ac.jp (S. Uchida), negishi@tufs.ac.jp (M. Negishi).

https://doi.org/10.1016/j.rmal.2025.100199
Received 21 November 2024; Received in revised form 9 March 2025; Accepted 9 March 2025
Available online 19 March 2025

classification of proficiency levels than the Common European Framework of Reference for Languages (CEFR), and the development and release of practical applications based on these standards. This study investigates the effectiveness of combining lexical metrics with AI-based scoring to assess CEFR-J levels in learner-generated English compositions, as implemented in the CEFR-based Writing Level Analyzer (CWLA).

The CEFR is a widely used language assessment framework categorizing language proficiency levels from A1 (beginner) to C2 (advanced) with level descriptors in the form of "can-do" statements for each skill. For instance, the self-assessment grid for B1-level writing specifies, "I can write simple connected text on topics which are familiar or of personal interest. I can write personal letters describing experiences and impressions" (Council of Europe, 2001). While this framework plays an essential role in multilingual settings such as Europe, its broad scale poses challenges for English as a Foreign Language (EFL) contexts such as Japan. To address this limitation, the CEFR-J framework was developed (Tono, 2019), providing a more fine-grained scale tailored to EFL learners, with 12 levels ranging from pre-A1 to C2. To the best of our knowledge, previous AES studies have not attempted alignment with CEFR-J levels, underscoring this study's originality.

AES systems generally rely on specific scoring criteria; however, these may not always be interpretable to users (accuracy-oriented models often use abstract metrics or deep-learning-based variables that lack transparency). Our approach combines straightforward lexical metrics with AI-based assessment anchored in the CEFR scale, which can generate feedback that is both meaningful and accessible to users. Furthermore, the AES system presented in this study prioritizes practical applicability and user-oriented interpretability, and it is currently accessible online under the name CWLA (https://cwla.langedu.jp/). We believe that moving beyond academic research to make such systems publicly accessible contributes to the outreach efforts in the field of English education.

The purpose of this study is to examine how AI scores and lexical metrics can be integrated within CWLA for accurate CEFR-J level assessment and to evaluate the accuracy of this approach compared to human ratings. To this end, this paper first reviews the latest research on AES using AI. It then introduces the data used to develop and validate CWLA, explains the lexical metrics and AI assessment methods employed by CWLA, and details the system's alignment with CEFR-J. Finally, the accuracy of CWLA is evaluated using two datasets, followed by a discussion and conclusion.

Related studies

Given the considerable interest in AES within English education, numerous studies have been conducted in this area. This section provides an overview, focusing particularly on recent research employing LLMs.

Evaluating the effectiveness of LLMs

Pack et al. (2024) investigated the effectiveness of four prominent LLMs—Google's PaLM 2, Anthropic's Claude 2, OpenAI's ChatGPT-3.5, and ChatGPT-4—in the automated scoring of essays written by English language learners. The study evaluated these models for cross-sectional and longitudinal validity and reliability, comparing their performance to human raters. The results revealed that ChatGPT-4 demonstrated the highest intra-rater reliability, with an intraclass correlation coefficient of 0.897 during the first scoring instance (T1) and 0.927 during the second scoring instance (T2), where T1 and T2 represent two separate scoring occasions with a minimum interval of 90 days. Its inter-rater reliability with human ratings was also the highest among the models tested, achieving a Pearson correlation of 0.731 at T1 and 0.638 at T2, indicating good validity despite a slight decline over time. While ChatGPT-4 showed promise in achieving reliable and valid automated essay scoring, the study identified some limitations. Specifically, ChatGPT-4 and other LLMs sometimes produced unexpected output, such as assigning continuous scores (e.g., 4.5 or 3.7) despite the ordinal rubric scale (1 to 6), or finishing students' incomplete sentences, which could artificially inflate scores.

Mizumoto et al. (2024) examined the efficacy of ChatGPT-4 in evaluating accuracy within the framework of complexity, accuracy, and fluency (CAF) in second-language (L2) acquisition. This study used the Cambridge Learner Corpus First Certificate in English (CLC FCE) dataset, an error-tagged learner corpus, to compare ChatGPT's performance in assessing errors or accuracy rates across 232 writing samples with those of human raters and Grammarly. The analysis revealed a moderate-to-strong correlation between ChatGPT's assessments and human accuracy assessments ($\rho = 0.79$), suggesting that ChatGPT can provide reasonably consistent scores in automated assessments. The study also compared Grammarly and ChatGPT and found that ChatGPT's assessments were closer to human judgments and students' writing scores. Specifically, ChatGPT showed a higher correlation with human ratings compared to Grammarly (ChatGPT $\rho = 0.79$ vs. Grammarly $\rho = 0.69$). Additionally, ChatGPT exhibited a stronger negative correlation with writing scores compared to Grammarly (ChatGPT $\rho = -0.63$ vs. Grammarly $\rho = -0.55$). The negative correlations indicate that a higher number of errors identified by a tool correlates with a lower writing score, and the stronger negative correlation for ChatGPT suggests its superior predictive validity. These findings suggest that ChatGPT aligns more closely with human expert judgments regarding linguistic accuracy and is more accurate than Grammarly in predicting overall writing quality. The study concluded that ChatGPT could serve as a valid tool for automated accuracy assessment in L2 writing. However, it also acknowledged certain limitations, such as the potential variability in ChatGPT's output, which must be considered.

Bannò et al. (2024) investigated the use of ChatGPT-4 for the analytic assessment of L2 writing proficiency in a zero-shot setting, where the model performs tasks without specific examples, aiming to provide more granular feedback than holistic scoring allows. The study primarily used the Write & Improve (W&I) dataset, annotated with holistic CEFR scores, and leveraged the larger EFCAMDAT dataset for pre-training a Longformer model (a transformer-based NLP model designed to process long documents efficiently using a combination of local and global attention mechanisms) to predict holistic scores. These predicted holistic scores, along with the original W&I holistic scores, were then fed to ChatGPT-4 to extract analytic components related to aspects such as grammatical

accuracy, vocabulary range, and coherence. Despite the lack of ground truth analytic scores, the study found that ChatGPT-4 appears to produce analytic scores reasonably related to the expected proficiency aspects. Specifically, the predicted scores for grammatical accuracy showed a correlation of 0.73 with grammatical error rate when using ground truth scores. For vocabulary range, the strongest correlation was 0.62 with the number of unique difficult words. These correlations suggest that ChatGPT-4′s predicted analytic scores capture relevant linguistic features, particularly for grammar and vocabulary.

These studies demonstrate the effectiveness of LLMs in AES. However, they do not directly assign CEFR levels, which may limit their applicability in contexts requiring precise alignment with CEFR standards.

Alignment with CEFR standards

Yancey et al. (2023) examined the ability of ChatGPT-4 to rate short L2 essays according to the CEFR scale. The study compared ChatGPT-4′s performance with that of ChatGPT-3.5, human raters, and modern Automated Writing Evaluation (AWE) methods. The results indicated that ChatGPT-4, when provided with calibration examples, could achieve a performance level comparable to that of current AWE systems. ChatGPT-4, when provided with only one calibration example per rating category, achieved a Quadratic Weighted Kappa (QWK) of 0.81, nearly matching the 0.84 QWK of the baseline AWE system. While performance did not improve significantly with additional calibration examples, agreement with human raters (measured by QWK) varied across test-takers′ L1s, ranging from 0.66 to 0.89. This calibration-based approach to guiding ChatGPT-4 with detailed prompts shows promise for automated CEFR level estimation.

Yamashita (2024) demonstrated the use of Many-Facet Rasch Measurement (MFRM; cf. Eckes, 2011) to evaluate ChatGPT-4 as an AES tool. Using data from the ICNALE corpus, the study compared ChatGPT-4′s scores on 136 essays with those assigned by 80 human raters. Both human raters and ChatGPT-4 differentiated essays across three CEFR proficiency levels. Correlations between human and ChatGPT-4 scores were moderate to strong ($r = 0.67$–0.82), with exact agreement around 50 %, and adjacent agreement (allowing a one-point difference) was considerably higher. Notably, ChatGPT-4′s ratings were found to be excessively consistent (overfit). The study also analyzed rater bias related to writer gender and found no significant bias from either human raters or ChatGPT-4. These findings suggest ChatGPT-4′s potential as an AES tool while highlighting MFRM's advantages as a comprehensive evaluation approach, offering insights beyond traditional correlation and agreement indices.

These studies are valuable contributions to examining the CEFR prediction capabilities of AI, providing evidence of its effectiveness to a certain extent. However, some research has also pointed out that LLMs struggle to accurately capture CEFR levels. Benedetto et al. (2025) examined LLM performance in classifying CEFR-based reading texts, automated essay scoring, and simulating learners at different proficiency levels. Their findings revealed that while LLMs possess some knowledge of CEFR, their outputs frequently exhibit inaccuracies, particularly a tendency to overpredict B2-level texts. Similarly, Uchida (2025) investigated the alignment of AI-generated texts with CEFR levels using ChatGPT-4o and a vocabulary level analyzer (CVLA; cf. Uchida & Negishi, 2018). The results indicated that ChatGPT-4o struggles to generate texts that consistently match the specified CEFR levels, often producing overly simplistic texts at lower levels and excessively complex texts at higher levels, along with notable topic biases. These findings suggest that direct methods for estimating CEFR or CEFR-J levels may not be effective, highlighting the need for alternative approaches.

Summary

In summary, recent research has shown a growing interest in using advanced LLMs, such as ChatGPT-4, for AES in L2 contexts. These studies have highlighted the capabilities of various models in evaluating aspects such as accuracy, linguistic proficiency, and overall writing quality, comparing them against traditional AES systems and human raters. Findings have indicated that ChatGPT-4 generally outperforms other LLMs in terms of reliability and validity. Studies such as those by Pack et al. (2024) and Mizumoto et al. (2024) showed that ChatGPT-4 can align well with human ratings, while others, such as that by Bannò et al. (2024), revealed ChatGPT-4′s potential in providing detailed analytic feedback on specific linguistic components. However, some challenges remain; for example, ChatGPT-4 occasionally demonstrates variability in scoring and may struggle to accurately align texts to proficiency levels. Yancey et al. (2023) and Yamashita (2024) further explored how AI models performed on CEFR scales, with Yancey's study demonstrating the benefits of calibration to improve ChatGPT-4′s CEFR alignment. Yamashita's use of MFRM offered a unique analytical perspective, suggesting that AI models such as ChatGPT-4 can serve as viable AES tools, albeit with certain discrepancies between them and human assessments. These studies shed new light on the potential of AI for CEFR alignment; however, none have progressed to the development of practical applications particularly for alignment with the CEFR-J scale.

Approach of this study

As outlined above, while existing research has focused on proficiency assessments using advanced AI and CEFR levels, to the best of our knowledge, no study to date has specifically addressed assessments based on CEFR-J levels. A primary reason for this gap is the lack of writing data (or other language skills) rated according to CEFR-J standards. One solution would be to create a dataset rated by CEFR-J levels; however, building a dataset of sufficient scale for machine-learning applications requires substantial time and effort. Therefore, this study adopts an alternative approach, using data with existing CEFR-based ratings to predict proficiency levels in a linear regression-like manner, aligning intermediary points with CEFR-J levels. This approach not only leverages readily available CEFR-tagged linguistic resources, but it also allows for straightforward level estimation, where increasing proficiency corresponds to rising indicator values—a model that is both simple and interpretable. To guide this study, the following research questions (RQs) were

formulated:

(1) How can AI scores and lexical metrics be integrated for accurate CEFR-J level assessments in learner writing?
(2) What is the accuracy of the proposed CWLA approach compared to human ratings for CEFR-J level assessments?

Additionally, many studies have focused solely on validating the accuracy of machine learning and AI techniques, and few have advanced to developing practical applications. By developing a web-accessible application that uses lexical metrics and AI-based assessments, this study contributes a tool that has a potentially substantial impact on both research and educational practice.

Datasets

Datasets tagged with CEFR levels are limited. For text corpora that serve as learner input (e.g., reading and listening materials), Uchida and Negishi (2018) analyzed text characteristics at each CEFR level by treating textbooks' overall level as equivalent to that of the individual passages they contain. However, textbooks contain passages of varying difficulty levels, ranging from simpler to more challenging content, and this dataset is not publicly available, which limits its applicability. In contrast, Arase et al. (2022) and Uchida et al. (2024) developed and released a dataset of approximately 17,000 sentences labeled with CEFR levels. The source texts, derived from resources such as Simple Wikipedia, are suitable for reading practice. These studies make valuable contributions by demonstrating the application of CEFR-tagged text data; however, none of them involve CEFR levels assigned to learner-generated writing, leaving a gap in the availability of such resources for AES research.

This study used two writing datasets: the ICNALE GRA Corpus and the CEFR ASAG Corpus (see the following subsections for details). The ICNALE GRA Corpus, while robust with human ratings, does not include CEFR levels. In contrast, the CEFR ASAG Corpus, although small in scale, includes CEFR levels assigned by multiple experts. The ICNALE GRA Corpus was employed to evaluate how closely the scores generated by our method align with human ratings. Meanwhile, the CEFR ASAG Corpus was used to map the scores of the proposed method from CEFR levels to CEFR-J levels, leveraging its proficiency labels for learner-generated texts. Together, these writing datasets provide a solid foundation for validating the effectiveness of the proposed method in terms of level alignment and scoring accuracy.

CEFR ASAG corpus

The CEFR ASAG Corpus, created by Tack et al. (2017), is a valuable resource annotated with CEFR levels for learner writing. In automated scoring, two primary systems are typically distinguished based on text length, type, and scoring method: Automated Essay Grading and Automated Short Answer Grading (ASAG). However, Tack et al. (2017), citing Burrows et al. (2015), highlighted that these distinctions can be ambiguous. Although this corpus was developed in the format of short answer grading, it is sufficiently versatile for broader applications. The tasks presented to learners in this study required open-ended responses of approximately 30–200 words, covering levels A1 to C2. This range aligns with the word counts often required for English composition in contexts such as Japanese university entrance exams, which typically range between 50 and 200 words, making it suitable for CEFR-J contexts.

Participants in Tack et al.'s study consisted of two groups: one group of learners enrolled in an e-learning platform and another group enrolled in university-level English language courses targeting specific CEFR levels. The CEFR ASAG Corpus provides proficiency assessments made by three certified Cambridge examiners, with each essay's level determined through majority voting among the examiners. Inter-rater reliability was high, with full agreement among the three raters for 44 % of the essays, and at least one matching pair of raters for 50 % of the texts, leaving only 6 % with complete disagreement. Consequently, the corpus included 299 essays with the following distribution across CEFR levels: A1 (18 texts), A2 (59 texts), B1 (113 texts), B2 (74 texts), C1 (30 texts), and C2 (5 texts). Owing to the small size of the CEFR ASAG Corpus, which is insufficient for statistical evaluation through data partitioning, the entire corpus was used for training in this study, while accuracy was evaluated through separate datasets and experiments.

ICNALE GRA corpus

The International Corpus Network of Asian Learners of English, Global Rating Archive (ICNALE GRA) is a learner corpus with annotated ratings, released in October 2023 as part of the ICNALE project, which collects and publicly shares English compositions by Asian learners (Ishikawa, 2020, 2023). This dataset includes 140 writing samples and 140 speaking (conversation) samples, both carefully sampled to maintain balance across the corpus. A unique feature of this dataset is that each sample has been rated by a pool of 80 raters who were carefully selected and trained for this task.

Each rater was tasked with two types of assessments. First, they provided a holistic score on a 100-point scale, estimating the overall quality of each essay or conversation. They then performed an analytical scoring across ten criteria: intelligibility, complexity, accuracy, fluency, comprehensibility, logicality, sophistication, purposefulness, willingness to communicate, and involvement, each rated on a 10-point scale. The average of these holistic and analytical scores constitutes the Overall Rating Score (ORS), which was used as a reference score in this study. By comparing the ORS with the scores generated through our method, the current study aimed to examine the correlation between human ratings and automated scoring.

Methodology

In this study, we combined AI-generated analytical scores with lexical metrics. Mizumoto and Eguchi (2023) asserted that while AI models, such as GPT-3, achieve a certain level of accuracy and reliability in AES, their performance improves significantly when integrated with traditional linguistic features. Their research demonstrated that incorporating elements such as lexical diversity and syntactic complexity into AES models enhances scoring accuracy. This suggests that although GPT-3 can provide valuable scoring independently, the most accurate results are achieved by combining AI with detailed linguistic metrics. This approach thus leverages the complementary strengths of AI and traditional statistical methods, resulting in a more robust framework for rating and providing feedback on student writing.

While some studies have reported that AI alone can achieve high scoring accuracy, relying solely on AI can reduce the interpretability of the scoring process. Following Mizumoto and Eguchi's (2023) findings, this study employed a hybrid method that combines AI-based analytical scores with conventional text features, balancing accuracy with interpretability in AES.

Calculation of lexical metrics

Numerous metrics have been proposed to assess the quality of learner writing, ranging from simple measures such as average sentence length to various metrics of lexical richness. Selecting metrics that are effective in combination with AI is challenging; therefore, this study adopted an approach based on three criteria: (1) the metric's proportional increase with proficiency level (to align regressionally with CEFR-J), (2) intuitive interpretability (to enable explicit feedback), and (3) ease of calculation (for implementation as a web application). Based on these criteria, this study utilized two primary metrics: the average vocabulary level (AvrDiff) and the proportion of B-level words to A-level words (BperA).

These metrics were originally proposed by Uchida and Negishi (2018). AvrDiff is calculated based on the CEFR-J Wordlist (https://www.cefr-j.org/download.html#cefrj_wordlist), where A1-level content words are assigned a value of 1, A2-level words a value of 2, B1-level content words a value of 3, and B2-level content words a value of 4, with the average then calculated. By contrast, BperA is calculated as the ratio of B-level content words to A-level content words (B content words / A content words).

Bulté and Housen (2012) highlighted the multidimensional nature of L2 complexity, categorizing lexical complexity into four dimensions: diversity, density, sophistication, and compositionality. The two metrics (AvrDiff and BperA) used in this study fall under lexical sophistication but target different aspects. AvrDiff reflects systemic lexical sophistication by measuring the average difficulty of vocabulary across a text, capturing the overall breadth and general quality of the learner's lexical repertoire. By contrast, BperA represents structural lexical sophistication by evaluating the ratio of B-level to A-level content words, focusing on the depth and progression within specific lexical complexities. According to Bulté and Housen (2012), systemic complexity assesses learners' linguistic resources on a broad scale, aligning with AvrDiff's role in evaluating vocabulary breadth, whereas structural complexity examines the refinement and advancement of specific linguistic elements, corresponding to BperA's focus on level-specific lexical development. By combining these complementary metrics, the study captures both the breadth and depth of lexical sophistication, thereby enabling a comprehensive assessment of learners' lexical complexity.

Piloting lexical metrics

A preliminary analysis was conducted to evaluate the applicability of these lexical metrics. This step allowed us to validate the reliability of these metrics in capturing linguistic features relevant to CEFR-based proficiency levels. Table 1 presents the results of our analysis of these metrics across the CEFR-based textbooks (adopted from Uchida & Negishi, 2018) and ASAG corpora. This table also includes additional two metrics: the Automated Readability Index (ARI) and VperSent, the average number of verbs per sentence used in Uchida and Negishi (2018) to show syntactic aspects of each corpus. These four metrics are validated as effective measures for determining the CEFR levels of reading texts in their study.

When reading and writing texts are compared, the ARI and VperSent metrics show significant differences. Although a direct increase with proficiency level is observed in writing, the values for these metrics deviate considerably from those in the textbook corpus, with VperSent values in writing generally higher than the textbooks—results that do not align intuitively. In contrast, AvrDiff and BperA display similar trends across both corpora, with values closely aligned by level, indicating a strong correlation between CEFR and vocabulary levels across both input (reading) and output (writing) texts. Although syntactic complexity metrics could theoretically complement these lexical measures, their application to learner data remains challenging. Learner-produced texts often contain grammatical errors that can affect part-of-speech tagging and dependency parsing, leading to some inconsistencies in syntactic complexity scores (cf. Huang et al., 2018). While recent advancements in NLP models trained with L2 data have improved parsing accuracy (Kyle & Eguchi, 2024), these improvements do not fully resolve the variability introduced by learner errors. Such technical challenges may limit the consistency of automatically calculated syntactic metrics, making lexical features a more stable alternative. Given these considerations, we prioritized the use of lexical metrics, which are both theoretically grounded and practically robust, to ensure the reliability and interpretability of the system's rating.

For the score calculations, we used Python's natural language processing library spaCy (version 3.7.2), employing the "en_core_web_sm" model as the dictionary base. Following Uchida and Negishi (2018), we assigned values of 1 for A1, 2 for A2, 3 for B1, and 4 for B2 levels. Using the values from the table above, we applied the following linear regression equations to convert the raw AvrDiff and BperA values to CEFR-aligned scores, adjusted to a maximum of 7. This final score, "voc_score," was then converted to a 10-point scale for use in writing assessment:

Table 1
Average lexical metrics of the CEFR-based textbooks and ASAG corpora.

	ARI		VperSent		AvrDiff		BperA	
	textbook	ASAG	textbook	ASAG	textbook	ASAG	textbook	ASAG
A1	5.73	2.27	1.49	2.07	1.31	1.29	0.08	0.08
A2	7.03	3.83	1.82	2.33	1.41	1.34	0.12	0.09
B1	10.00	5.87	2.37	3.08	1.57	1.47	0.18	0.15
B2	12.33	9.37	2.88	3.79	1.71	1.67	0.26	0.26

$$\text{score\_BperA} = \text{BperA} \times 14.04 + 0.65$$

$$\text{score\_AvrDiff} = \text{AvrDiff} \times 7.56 - 8.30$$

$$\text{voc\_score} = \left(\frac{\min(\text{score\_BperA}, 7) + \min(\text{score\_AvrDiff}, 7)}{2}\right) \times \frac{10}{7}$$

AI scoring

For AI-based scoring, this study calculated a numerical score aligned with CEFR-J levels rather than directly classifying texts into CEFR levels. Uchida (2024), using the ICNALE GRA dataset, performed both holistic scoring (rated out of 100) and analytical scoring (10 criteria rated out of 10), demonstrating that analytical scoring produced superior results. Specifically, the correlation was 0.828 for holistic scoring with ChatGPT-4, compared to 0.873 for analytical scoring; moreover, holistic scoring often yielded identical scores, resulting in skewed scatter plot distributions. Additionally, Bannò et al. (2024) demonstrated that analytical scoring using ChatGPT-4 showed significant correlations across various CEFR-based aspects of L2 writing, with particularly high reliability in grammar and vocabulary evaluation. Based on these findings, this study also adopted analytical scoring. However, recognizing that the 10 criteria used in ICNALE GRA may be overly granular and conceptually redundant, we streamlined our scoring to five key criteria: Complexity, Accuracy, Fluency, Logicality, and Sophistication. The first three criteria are commonly known as CAF measures (cf. Wolfe-Quintero et al., 1998) and are traditional indicators of linguistic features in writing evaluation. Previous research has found these CAF measures to be effectively scored using AI (Mizumoto et al., 2024; Pfau et al., 2023). The latter two criteria assess semantic aspects that are difficult to quantify with conventional measures. While further investigation is needed to confirm that each score genuinely represents its respective criterion, this study focused on aggregating these criteria to a total score of 50, which serves as the AI-generated score.

We employed the ChatGPT-4o model (gpt-4o-2024–08–06) for AI scoring, with the randomness parameter (temperature) set to 0 to ensure consistent results. The prompt used was as follows:

> You are an experienced English teacher. Your task is to rate a passage written by an English learner on a 10-point scale in terms of the following points: Complexity, Accuracy, Fluency, Logicality, and Sophistication. The output should be in the JSON format with the keys "Complexity," "Accuracy," "Fluency," "Logicality," and "Sophistication."

By using AI to assign numerical scores, these scores can be aligned with the finer-grained CEFR-J levels within the broader CEFR framework. If this task were framed as a classification problem—predicting CEFR levels as discrete labels—the AI's accuracy would likely be lower owing to its limited specific knowledge of CEFR-J level distinctions. To support this assumption, we conducted an experiment using the ASAG corpus, in which ChatGPT-4o (gpt-4o-2024–08–06) was tasked with predicting CEFR levels.

The results in Table 2 suggest that the match rate between AI predictions and human-assigned CEFR levels was 175 out of 299, or 58.53 %. While many predictions fell within adjacent levels, notable discrepancies appeared at certain levels, particularly at B1, B2, C1, and C2, where AI's scores diverged significantly from those of human ratings. These findings highlight limitations in the AI's ability to accurately classify even on the broader CEFR scale, let alone the more detailed CEFR-J scale. To address this issue, this study adopted a regression-based approach instead of a classification model for CEFR level alignment. This approach enables more accurate scoring by capturing proficiency gradations more effectively, particularly within the finer subdivisions of the CEFR-J framework.

Table 2
ASAG Corpus levels and ChatGPT-4o estimations.

ASAG/ChatGPT	A1	A2	B1	B2	C1	total
A1	**12**	6				18
A2	4	**55**				59
B1		43	**67**	3		113
B2		1	32	**40**	1	74
C1			4	25	**1**	30
C2				5		5
total	16	105	103	73	2	299

Table 3
Statistics of CEFR levels and CWLA scores in the ASAG Corpus.

	average	min	max	SD
A1	36.07	24.46	45.24	5.48
A2	40.85	30.04	54.56	6.19
B1	50.08	33.46	65.90	6.10
B2	60.75	45.69	80.00	6.30
C1	66.64	52.78	77.98	6.61
C2	72.19	66.59	75.00	3.27

Table 4
Alignment of CWLA scores with CEFR-J levels.

level	min	max	range
PreA1	0	25	25
A1.1	26	29	3
A1.2	30	33	3
A1.3	34	37	3
A2.1	38	41	3
A2.2	42	45	3
B1.1	46	50	4
B1.2	51	55	4
B2.1	56	60	4
B2.2	61	66	5
C1	67	72	5
C2	73	100	27

Alignment to CEFR-J levels

This study's alignment to CEFR-J levels is based on the analysis of the CEFR ASAG Corpus, which is annotated with CEFR levels, using CWLA scores. As mentioned earlier, no existing corpus is annotated with CEFR-J levels; therefore, the alignment is achieved by subdividing the average scores for each CEFR level.

In the CWLA system, the final score was derived by combining the 10-point voc_score and 50-point AI score, then scaling the result to a 100-point scale. The calculation formula was as follows:

$$\text{CWLA\_score} = \left\lfloor \frac{\text{voc\_score} + (\text{Complexity} + \text{Accuracy} + \text{Fluency} + \text{Logicality} + \text{Sophistication})}{60} \times 100 \right\rfloor$$

The balance between the vocabulary score (10 points) and the AI score (50 points) was determined empirically. Prior research has indicated that AI achieves high accuracy in scoring; thus, a greater weight was allocated to the AI component to fully leverage its scoring capabilities. This weighting reflects the relative strength of AI-based assessments while allowing the lexical metrics to provide complementary insights into learner proficiency.

Using the formula above, we calculated the CWLA_score for each CEFR level in the ASAG corpus, which contains learner writing samples categorized by CEFR levels. Table 3 summarizes the average CWLA_score for each proficiency level.

Based on these results, we assigned ranges to each CEFR-J level, as shown in Table 4. This allocation assumes that the finer subdivisions of CEFR-J represent intermediate points within each CEFR level, leading to relatively regular intervals between levels. Decimal points are truncated in the final scores. While this alignment is somewhat subjective, raters who are well versed in CEFR-J levels assess its validity in the following sections.

Validation of CEFR-J alignment

The CWLA is validated using two methods. The first involves using the ICNALE GRA dataset, which contains reliable human-assigned scores, to examine how closely CWLA scores align with human ratings. This step ensures the overall reliability of the scores generated by CWLA. The second method involves expert judgment, aimed at verifying the validity of CEFR-J level subdivisions. As these subdivisions were created mechanically, this step is essential to confirm their appropriateness.

Validation using ICNALE GRA

To validate CWLA's scoring accuracy, we used the ICNALE GRA dataset, which includes 140 writing samples rated by human raters. Each sample has an Overall Rating Score (ORS), calculated as the average of holistic and analytical scores, providing a reliable reference for comparison. This validation assesses how closely CWLA scores align with human ratings in the ICNALE GRA and includes a comparison with scores generated by ChatGPT alone, scaled to 100 points for consistency.

Fig. 1 presents a scatter plot of ChatGPT-4o's scores (doubled for a 100-point scale) versus the ORS. The correlation coefficient

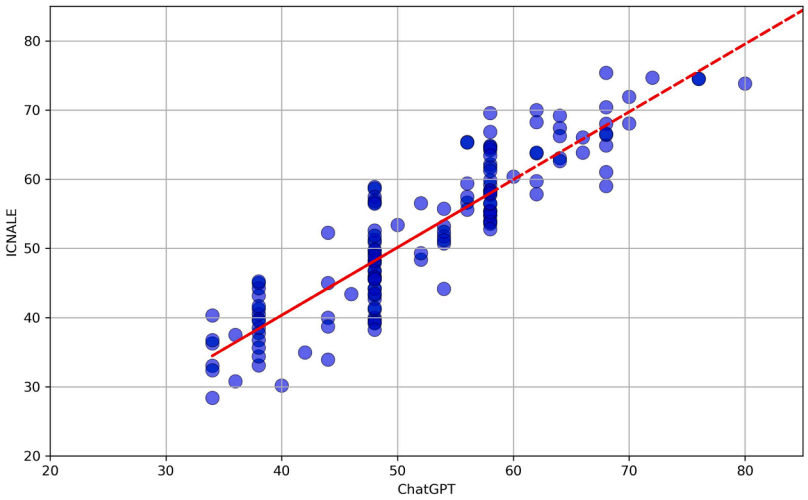

Fig. 1. Scatter plot of ChatGPT scores vs. ICNALE GRA scores.

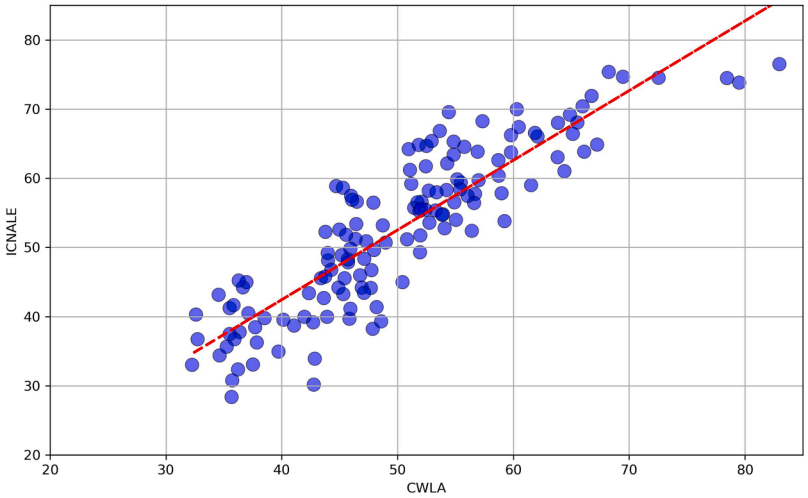

Fig. 2. Scatter plot of CWLA scores vs. ICNALE GRA scores.

between ChatGPT scores and the ORS, computed using Python's scipy.stats, is 0.90 ($p < .0001$, $R^2 = 0.81$), showing a high level of correlation. However, the scatter plot reveals a tendency for ChatGPT to assign identical scores, which diverges from the variability observed in human ratings. This tendency was also noted by Yamashita (2024), who highlighted ChatGPT's over-consistent scoring behavior, often assigning scores within a narrow range and avoiding extreme values, which contrasts with the broader distribution typically observed in human raters' assessments.

Fig. 2 displays the scatterplot comparing CWLA scores to the ORS, with a correlation coefficient of 0.88 ($p < .0001$, $R^2 = 0.77$). While this correlation is slightly lower than that of ChatGPT's, CWLA scores exhibit a wider distribution of values. This broader spread suggests that CWLA's scoring approach aligns more closely with human judgment, capturing the variability that human raters tend to display in their assessments.

To further examine this variability, we conducted entropy calculations as a measure of data dispersion. Entropy is calculated using the following formula:

$$H = -\sum_{i=1}^{n} P(x_i)\log_2 P(x_i)$$

where $p(x_i)$ represents the probability of each unique scorex_i, quantifies the level of unpredictability or diversity in a set of values. Higher entropy indicates greater variability. For example, if the scores are spread out evenly across all possible categories, the entropy will be higher than when most scores cluster around just a few categories. Using the entropy function from scipy.stats (with bins=25), ChatGPT scores yielded an entropy of 3.52, indicating a narrower score range and less diversity. By contrast, CWLA scores achieved an

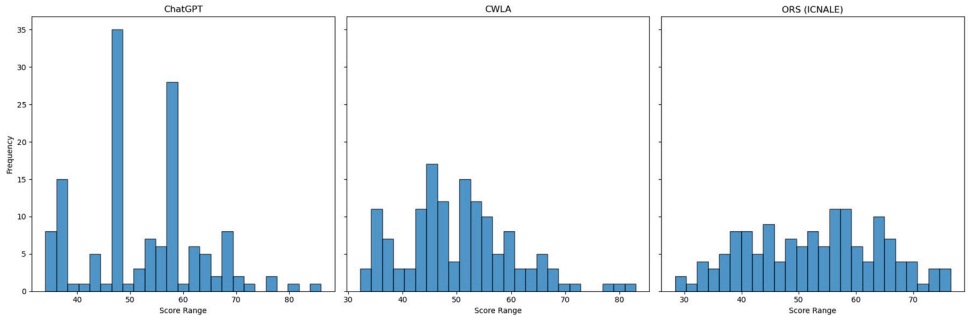

Fig. 3. Histograms of each score distribution.

entropy of 4.08, more closely aligning with the ICNALE ORS entropy of 4.45. Furthermore, as shown in Fig. 3, ChatGPT's scores tend to concentrate on specific values. These findings suggest that CWLA not only approximates the mean scores of human raters but also better replicates the variability observed in human scoring.

Validation by English education experts

Next, we assessed whether CWLA's CEFR-J level estimations aligned with expert judgments. For this validation, three experts proficient in CEFR and CEFR-J rated the levels assigned by CWLA to determine their accuracy. The evaluation was conducted independently by each expert using a Word document, with CEFR-J writing level descriptors provided as reference.

Two of the experts are native English speakers with over 15 years of experience teaching English at Japanese universities, which makes them highly familiar with Japanese students' proficiency levels and the CEFR(-J) framework. The third expert is a native Japanese graduate student specializing in English education, with over three years of high school teaching experience. This individual also works in quality management at a company offering English revision services for Japanese high school students and possesses a thorough understanding of CEFR(-J).

For this validation, we selected writing samples from two groups to cover both low and high proficiency levels: (1) letters written by Japanese junior high school students about weekend activities, and (2) essays by Japanese high school students on whether parents should control internet use (see Appendices A1 and A2 for details). We gathered three samples per level from pre-A1 to B2.2, which resulted in 30 writing samples. Each rater reviewed these samples, yielding 90 evaluation data points (3 raters $\times$ 30 samples). While these writing samples cannot be widely shared owing to the lack of explicit consent for public distribution at the time of collection, we aim to build a CEFR-J annotated writing corpus in the future, incorporating CWLA results with expert-assigned CEFR-J levels. Using Scikit-learn (version 1.5.1), we calculated the inter-rater Cohen's kappa scores, which were 0.59, 0.67, and 0.85, indicating moderate to substantial agreement between individual raters. The overall Fleiss' kappa for the three raters was 0.70, demonstrating a considerable level of agreement among raters. These kappa values suggest that the raters maintained a consistent standard in their CEFR-J level assessments, supporting the reliability of the human reference data used in this study.

Table 5 summarizes the comparison between CWLA-assigned levels (rows) and the ratings from the three raters (columns). The results showed a high agreement rate of 83.33 % (75 out of 90 cases). While CWLA showed a slight tendency to assign higher scores at the B2.2 level, all discrepancies fell within adjacent levels, which indicated no major deviations. These results suggest that CWLA can provide reliable CEFR-J level estimations that closely align with expert judgments, especially within practical educational contexts.

Table 5
Cross tabulation of CWLA level assignments (row) and raters' judgments (column).

level/judge	preA1	A1.1	A1.2	A1.3	A2.1	A2.2	B1.1	B1.2	B2.1	B2.2	total
preA1	**8**	1									9
A1.1		**8**	1								9
A1.2		1	**7**	1							9
A1.3				**7**	2						9
A2.1					**9**						9
A2.2						**9**					9
B1.1						1	**8**				9
B1.2							1	**8**			9
B2.1								2	**7**		9
B2.2									5	**4**	9
total	8	10	8	8	11	10	9	10	12	4	90

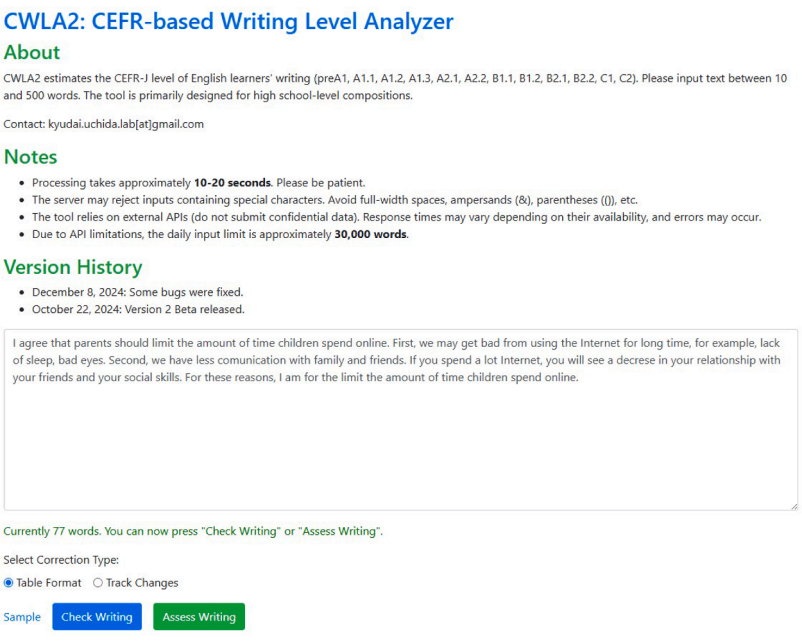

Fig. 4. The front page of CWLA.

The CWLA web application

The findings above suggest that the CWLA's CEFR-J level assessment achieved a reliable level of accuracy. To make this accessible to a broader audience, we developed a web application using Python's Flask framework (https://cwla.langedu.jp/).

The user interface was designed to be simple and intuitive. As shown in Fig. 4, the main screen consists of an input box, a selection for feedback format, and two buttons: "Check Writing" and "Assess Writing." Users paste their English text into the input box, select a feedback format (either "table format" or "track changes"), and then click "Assess Writing" to see the results, as displayed in Fig. 5. The output screen provides color-coded words based on the CEFR-J Wordlist, a detailed breakdown of scores, and visual representations of these scores. This summary is followed by a revision of the input text (not shown in Fig. 5), with suggested edits according to the

Fig. 5. A sample analysis of CWLA.

selected feedback format. Additionally, when users click “Check Writing” on the top screen, indirect feedback on the entered text is displayed, offering guidance for corrections. Examples of this type of feedback include comments such as the following:

> **First, we may get bad from using the Internet for long time, for example, insomnia, poor eyesight**: The phrase "get bad" is unnatural. Consider using a phrase like "experience negative effects" or "suffer from issues." Also, consider adding an article before "long time."

This message was generated using ChatGPT-4o-mini with the following prompt:

> The following English composition was written by a learner. Please provide advice for improvement in English. Only give advice without showing explicit corrections (feedback should be indirect with some clues). No advice is needed for sentences that do not require any corrections.

It should be noted that this prompt is a beta version and may undergo further improvements in the future. The mini version of ChatGPT-4o is utilized to ensure faster response times, and the system is designed to automatically select the latest version available.

Thus, CWLA is a simple, user-friendly tool accessible to anyone, with feedback and revision features that support self-directed learning. With these features, CWLA can be effectively used to encourage autonomous learning and help learners gain insight into their development in language skills.

Discussion

This study has introduced the development process and accuracy evaluation of CWLA, a web-based application designed to measure the CEFR-J levels of English learners’ writing. The findings are summarized below along with key discussion points in response to the research questions: (1) “How can AI scores and lexical metrics be integrated for accurate CEFR-J level assessments in learner writing?” and (2) “What is the accuracy of the proposed CWLA approach compared to human ratings for CEFR-J level assessments?”

Integrating AI scores and lexical metrics for CEFR-J level assessment

As demonstrated in numerous previous studies, contemporary generative AI systems exhibit high accuracy in rating and correcting writing. The approach of combining AI with existing lexical metrics, as indicated in Mizumoto and Eguchi (2023), offers the advantage of avoiding the complete black-boxing of AI assessment processes. In this study, we adopted two stable lexical metrics—average lexical difficulty (AvrDiff) and the ratio of CEFR B-level content words to A-level content words (BperA)—as they provided consistent values across both textbook and writing corpora.

To enable multidimensional rating, the CWLA system uses ChatGPT-4o to calculate scores based on the following five perspectives: “Complexity,” “Accuracy,” “Fluency,” “Logicality,” and “Sophistication.” While the validity of these dimensions in effectively capturing their respective aspects requires further investigation, previous research (e.g., Bannò et al., 2024) suggests that AI-based ratings across multiple textual dimensions can achieve a certain level of reliability. To align with the CEFR-J scale, we employed a method that subdivides levels using midpoint values derived from average scores for each level in a writing corpus annotated with CEFR levels. This process involves subjective elements in balancing the AI score with the vocabulary score and determining the range of CEFR-J levels. Therefore, the validity of these aspects was examined under the second research question.

Accuracy of CWLA compared to human ratings

The accuracy of CWLA was evaluated using two approaches. First, the degree of agreement between CWLA scores and human ratings was examined using the ICNALE GRA dataset, which contains 140 essays rated by 80 human raters. The correlation coefficient between CWLA’s overall scores and human ratings was found to be 0.88, indicating high reliability. From the perspective of entropy, it demonstrated behavior similar to that of human raters. Second, the functionality of the CEFR-J conversion was tested. Ratings by three experts for 30 essays yielded an agreement rate of 83.33 %, further supporting the system’s effectiveness.

A key point for discussion is whether the combination of AvrDiff and BperA represents the optimal lexical metrics for integration with AI. While prior research has shown that AI alone can achieve high accuracy, the primary rationale for adopting these two metrics in this study was their interpretability for users and their stability for computational rating. However, traditional AES studies have explored various textual features, including lexical diversity, complexity, and syntactic metrics. Identifying more effective combinations to enhance accuracy remains an area for future research.

Educational implications

Beyond its technical accuracy, the practical impact of CWLA in educational settings remains an important consideration. Numerous studies have demonstrated the effectiveness of AI in English writing (for a detailed review, see Teng, 2024a; Pratama & Sulistiyo, 2024). For instance, Teng (2024b) highlighted the impact of ChatGPT on EFL writing by improving students’ motivation, self-efficacy, and engagement through immediate and detailed feedback. Students appreciated the tool’s ability to enhance grammar, structure, and clarity, suggesting its potential as a valuable resource for writing development in educational contexts. Similar effectiveness can reasonably be expected from CWLA. However, this claim remains speculative at present. To address this, we are currently developing a

system capable of managing individual learning progress and conducting a pilot study in Japanese high schools. A preliminary survey conducted with 39 third-year Japanese high school students revealed the following results: students rated the usefulness of the pre-check feature at 3.98 out of 5 (SD = 1.07) on a five-point scale, the usefulness of English corrections at 4.36 out of 5 (SD = 1.13), and the fairness of scores at 4.00 out of 5 (SD = 1.13). These findings indicate that learners hold a generally positive perception of CWLA. The details of this experiment will be reported in a future publication.

Conclusion

This study showed that CWLA scores, based on lexical metrics and AI assessment, had a correlation of 0.88 with ICNALE GRA scores. Furthermore, the system's conversion of scores to CEFR-J levels, based on a CEFR-aligned writing corpus, achieved an 83.33 % match with expert judgments, validating the CWLA system's alignment with human evaluative standards. To the best of our knowledge, this development represents the world's first web application capable of accurately determining CEFR-J levels. This makes it noteworthy not only for its accessibility but also as a practical contribution of academic work to the broader public.

Reliable automated scoring opens up new avenues for research. For instance, the ability to classify writing data by CEFR-J levels with high accuracy can help reveal behaviors specific to each proficiency level. As a readily accessible tool, CWLA also promotes self-directed learning, allowing for studies on learner behavior and the educational effectiveness of AI-assisted feedback. CWLA not only advances AES research but also offers tangible benefits for educators and learners.

One limitation of this study is that it did not consider the impact of topics on writing performance. Previous studies, such as Yang et al. (2015), have indicated a relationship between topic and learner output. While CWLA successfully captures general writing trends, incorporating topic-specific scoring could further enhance accuracy, which is potentially achievable through AI-driven topic adaptation. Another consideration is that validation was conducted solely on writing samples from Japanese middle and high school students. Given the focus on CEFR-J alignment, this was a natural choice; however, to increase generalizability, future research could involve learners from diverse backgrounds. Finally, we must acknowledge that, despite efforts to minimize randomness, reproducibility remains a challenge when using AI. Owing to the nature of LLMs, results may vary across model updates, which could affect consistency. Users of our application should be informed of these limitations to set appropriate expectations regarding the stability of the results.

Use of AIs

This study examined the capability of ChatGPT to assess learners' writing and thus extensively used it for the experiments and construction of the web application. Additionally, during the preparation of this work, the authors used ChatGPT-4o to improve, proofread, and translate the manuscript. After using this tool, the authors reviewed and edited the content as required, and they take full responsibility for the content of the publication.

CRediT authorship contribution statement

Satoru Uchida: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Masashi Negishi**: Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Satoru Uchida reports financial support was provided by Japan Society for the Promotion of Science. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP20H00095 and JP23K21949. We would like to express our sincere gratitude to the anonymous reviewers for their detailed and insightful comments, which greatly contributed to improving the quality of this manuscript. Any remaining errors or oversights are solely the responsibility of the authors.

(A.1)

About a weekend

Your friend Nancy, who lives in the United States, is researching how Japanese students spend their weekends. She sent an email asking about what you did last Sunday. She wants to know where you went, what you did, and your thoughts about it. Continue writing a reply to her email, using a dictionary if needed, following the opening provided below.

Dear Nancy,

(Last Sunday,)

(A.2)

The use of the Internet by children

Write your opinion on the following question and provide two reasons for your answer in 80 to 100 words (The use of a dictionary is not prohibited).

Do you agree that parents should limit the amount of time their children spend online?

Appendices

Prompts used in the writing tasks for the CWLA validation

References

Arase, Y., Uchida, S., & Kajiwara, T. (2022). CEFR-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 6206–6219). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.416.

Bannò, S., Vydana, H. K., Knill, K., & Gales, M. (2024). Can GPT-4 do L2 analytic assessment?. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 149–164). Association for Computational Linguistics.

Benedetto, L., Gaudeau, G., Caines, A., & Buttery, P. (2025). Assessing how accurately large language models encode and apply the common European framework of reference for languages. *Computers and Education: Artificial Intelligence, 8*, Article 100353. https://doi.org/10.1016/j.caeai.2024.100353

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in sla* (pp. 21–46). John Benjamins. https://doi.org/10.1075/lllt.32.

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education, 25*(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8

Council of Europe. (2001). *Common european framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

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, 29*, 14151–14203. https://doi.org/10.1007/s10639-023-12402-3

Eckes, T. (2011). *Introduction to many-facet rasch measurement: Analyzing and evaluating rater-mediated assessments*. Peter Lang.

Huang, Y., Murakami, A., Alexopoulou, T., & Korhonen, A. (2018). Dependency parsing of learner English. *International Journal of Corpus Linguistics, 23*(1), 28–54.

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

Ishikawa, S. (2020). Aim of the ICNALE GRA project: Global collaboration to collect ratings of Asian learners' L2 English essays and speeches from an ELF perspective. *Learner Corpus Studies in Asia and the World, 5*, 121–144. https://doi.org/10.24546/81012494

Ishikawa, S. (2023). *The ICNALE guide: An introduction to a learner corpus study on Asian learners' L2 English*. Routledge.

Kyle, K., & Eguchi, M. (2024). Evaluating NLP models with written and spoken L2 samples. *Research Methods in Applied Linguistics, 3*(2), Article 100120. https://doi.org/10.1016/j.rmal.2024.100120

Lim, C. T., Bong, C. H., Wong, W. S., & Lee, N. K. (2021). A comprehensive review of automated Essay Scoring (AES) research and development. *Pertanika Journal of Science & Technology, 29*(3). https://doi.org/10.47836/pjst.29.3.27

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246–276. https://doi.org/10.1191/0265532202lt230oa

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), Article 100050. https://doi.org/10.1016/j.rmal.2023.100050

Mizumoto, A., Shintani, N., Sasaki, M., & Teng, M. F. (2024). Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment. *Research Methods in Applied Linguistics, 3*(2), Article 100116. https://doi.org/10.1016/j.rmal.2024.100116

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

Pratama, A., & Sulistiyo, U. (2024). A systematic review of artificial intelligence in enhancing English foreign learners' writing skill. *PPSDP International Journal of Education, 3*(2), 170–181. https://doi.org/10.59175/pijed.v3i2.299

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), Article 100083. https://doi.org/10.1016/j.rmal.2023.100083

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

Tack, A., François, T., Roekhaut, S., & Fairon, C. (2017). Human and automated CEFR-based grading of short answers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 169–179). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5018.

Teng, M. F. (2024a). A systematic review of ChatGPT for English as a Foreign Language writing: Opportunities, challenges, and recommendations. *International Journal of TESOL Studies, 6*(3), 36–57. https://doi.org/10.58304/ijts.20240304

Teng, M. F. (2024b). ChatGPT is the companion, not enemies: EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. *Computers and Education: Artificial Intelligence, 7*, Article 100270. https://doi.org/10.1016/j.caeai.2024.100270

Tono, Y. (2019). Coming full circle: From CEFR to CEFR-J and back. *CEFR Journal - Research and Practice, 1*, 5–17. https://doi.org/10.37546/jaltsig.cefr1-1

Uchida, S. (2024). Evaluating the accuracy of ChatGPT in assessing writing and speaking: A verification study using ICNALE GRA. *Learner Corpus Studies in Asia and the World, 6*, 1–12. https://doi.org/10.24546/0100487710

Uchida, S. (2025). Generative AI and CEFR levels: Evaluating the accuracy of text generation with ChatGPT-4o through textual features. *Vocabulary Learning and Instruction, 14*(1), 2078. https://doi.org/10.29140/vli.v14n1.2078

Uchida, S., Arase, Y., & Kajiwara, T. (2024). Profiling English sentences based on CEFR levels. *ITL - International Journal of Applied Linguistics, 175*(1), 103–126. https://doi.org/10.1075/itl.22018.uch

Uchida, S., & Negishi, M. (2018). Assigning CEFR-J levels to English texts based on textual features. In *, 4. Proceedings of Asia Pacific Corpus Linguistics Conference* (pp. 463–467). Asia Pacific Corpus Linguistics Conference.

Uto, M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika, 48*(2), 459–484. https://doi.org/10.1007/s41237-021-00142-y

Wolfe-Quintero, K., Inagaki, S., & Hae-Young, K. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. University of Hawai'i Press.

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), Article 100133. https://doi.org/10.1016/j.rmal.2024.100133

Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 576–584). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.49.

Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing, 28*, 53–67. https://doi.org/10.1016/j.jslw.2015.02.002