Research Memorandum
ETS RM–24-05

Exploring the Provision of Artificial Intelligence–Based Feedback for TOEFL Junior® Writing Practice Tasks

Mikyung Kim Wolf
Michael Suhan
Jean Alderman

April 2024

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Exploring the Provision of Artificial Intelligence–Based Feedback for TOEFL Junior® Writing Practice Tasks

Mikyung Kim Wolf, Michael Suhan, and Jean Alderman
ETS, Princeton, New Jersey, United States

April 2024

Corresponding author: Mikyung Kim Wolf, E-mail: mkwolf@ets.org

Suggested citation: Wolf, M. K., Suhan, M., & Alderman, J. (2024). *Exploring the provision of artificial intelligence–based feedback for TOEFL Junior® Writing practice tasks* (Research Memorandum No. RM-24-05). ETS.

Action Editor: Jonathan Schmidgall

Reviewers: Ching-Ni Hsieh and Lorraine Sova

Abstract

This research memorandum reports on an exploratory study of developing an artificial intelligence (AI)–based feedback prototype tool and examining its usability for potential TOEFL Junior® Writing test users. The recently launched TOEFL Junior Writing test employs ETS's automated writing evaluation (AWE) tool to score test takers' responses. Utilizing the ETS's AWE engine and an existing database, we designed an approach to feedback that was closely linked to the key writing dimensions and proficiency descriptors as outlined in the TOEFL Junior Writing rubrics. A small-scale usability study was conducted with 14 students and seven teachers from South Korea and Türkiye. The participants perceived the usefulness of the feedback tool positively, but they also provided practical suggestions to further improve the tool. The promises and limitations of the tool are discussed.

Keywords: artificial intelligence (AI), writing feedback, automated writing evaluation (AWE), English as a foreign language (EFL), EFL writing, TOEFL Junior tests

Acknowledgments

The authors wish to acknowledge the following project team members who provided technological expertise in building the prototype tool: Michael Wagner, Paul Brost, Ramin Maxwell Hemat, Nathan Lederer, and Stephanie Dais. The authors also extend their gratitude to the staff from ETS Global Channel, Goal Testing in Türkiye, and Metacurio in Korea for their assistance with participant recruitment as well as to the teachers and students for their participation in this project.

Recently, ETS launched the TOEFL Junior® Writing test in response to a growing demand for a four-skills assessment for young language learners worldwide. Combined with the existing listening, reading, and speaking tests, the TOEFL Junior tests now offers a complete four-skills assessment. This TOEFL Junior Writing test evolved from the writing section of the previous TOEFL Junior Comprehensive test, which was discontinued in 2017. A major change from the previous test lies in the use of an automated writing evaluation (AWE) engine for scoring writing responses, while maintaining the same task types. The advancements in technology and artificial intelligence (AI) capabilities increasingly enable the generation of a high-performing AWE engine that accurately predicts human scoring.

The adoption of an AWE engine in the TOEFL Junior Writing test provides an opportunity to offer AI-based feedback to improve students' writing. Previous research has demonstrated a number of advantages of automated feedback on students' writing such as immediateness, personalization, specificity, error correction, consistency, and labor and cost effectiveness (e.g., Fu et al., 2024; Hoang & Kunnan, 2016; Ranalli et al., 2017; Wilson et al., 2021). Research has shown mixed findings on the effectiveness of automated feedback on learning outcomes depending on the quality and types of feedback as well as students' background. However, as Hattie and Timperley (2007) noted, feedback is essential to learning and achievement.

In light of this, the present study aimed to explore the use of ETS's AWE engine to provide AI-based feedback for the TOEFL Junior Writing test that is timely and beneficial for teachers and students. The provision of useful feedback for students' writing can also enhance the utility of the TOEFL Junior tests, as one of their intended uses was to guide teaching and learning (So et al., 2015). This study largely involved two phases: (a) developing an automated feedback prototype tool and (b) examining its usability. During the prototype development, we explored appropriate types of AI-based feedback for target users derived from ETS's AWE engine. Then, we gathered users' perceptions about the usability and usefulness of the feedback prototype tool. We anticipated that the findings of the study would offer useful insights into the types of feedback and the ways to deliver it to users. For instance, AI-based feedback could be incorporated into a score report. Additionally, we expected the findings

would inform further refinement of the feedback tool, particularly in situations where AI-based feedback could be presented as a standalone tool rather than integrated into the score report.

In this report, we first provide a brief overview of the TOEFL Junior Writing test, focusing on its task types, the scoring rubric that covers key subskills assessed, and the AWE engine features. The subsequent section describes the development process of the feedback prototype tool and our approaches to determining feedback areas. We then report on the usability study and its findings. This report closes with a discussion on the implications for future development and research directions in providing AI-based feedback to potential TOEFL Junior Writing test users.

TOEFL Junior Writing Tests

The TOEFL Junior tests comprise three tests: (a) the TOEFL Junior Standard test, consisting of Listening, Language Form and Meaning, and Reading sections, (b) the TOEFL Junior Speaking test, and (c) TOEFL Junior Writing test. The tests are provided as standalone modules intended for students learning English as an additional, second, or foreign language (EAL/ESL/EFL) in secondary schools. The tests are intended to assess students' English proficiency as needed to participate in English-medium instructional settings. The tests are designed to measure students' English proficiency against an international benchmark such as the Common European Framework of Reference for Languages (CEFR), track students' progress, inform placement decisions in English language programs, and provide information that can support instruction (ETS, 2019; So et al., 2015).

As mentioned earlier, the TOEFL Junior Writing test was part of the former TOEFL Junior Comprehensive test and was reintroduced in August 2023 as a standalone module. It measures test takers' computer-based English writing ability to communicate within interpersonal, school navigational, and academic domains across four task types (Edit, E-mail, Opinion, and Listen-Write). Table 1 provides a description of each task type of the TOEFL Junior Writing test.

Table 1. Description of TOEFL Junior Writing Test Tasks

Task	Description	Response format	Allotted time in minutes
Edit	Test takers select the correct option to complete sentences in a paragraph about an academic topic.	MC	2.5
E-mail	Test takers write a reply to an e-mail about a school-related topic to answer questions from an e-mail.	CR	7
Opinion	Test takers write a response to a question or statement about a specific topic, expressing and providing support for their opinion about the topic.	CR	10
Listen-Write	Test takers listen to a talk about an academic topic supported by visuals and then summarize the main points of the talk. They may take notes while listening and use the notes to support the response.	CR	10

Note. MC = multiple choice; CR = constructed response.

The three CR tasks are scored by ETS's AWE engine, which is trained on human ratings using a separate holistic rubric for each task. Because each task type entails different genres, task-specific rubrics are utilized with a 5-point score scale (0 to 4). Descriptors on all three rubrics cover the dimensions of content, organization (particularly, cohesion and coherence), and language use as key writing subskills. Appendix A displays the descriptor of the highest score point of each task rubric (see the complete rubrics at https://www.ets.org/pdfs/toefl/toefl-junior-writing-scoring-guide.pdf). The features included in the AWE scoring models are conceptually aligned with the descriptors from the scoring rubric. Table 2 presents a summary of the major AWE engine features, mapped to the subskill dimensions in the rubric. Note that microlevel features are not included in the table.

Table 2. Major Feature Categories Included in the Scoring Engine by Rubric Dimensions

Rubric skill dimension	Scoring engine feature categories
Content	Similarity of response to prompt Detailing features Use of subjective language
Cohesion and coherence	Use of pronouns Use of discourse cues Connectedness of ideas
Language use	Accuracy of lexis, syntax, morphology, and mechanics Variety of lexis and syntax Complexity of lexis and syntax

Note. Detailing features, as listed under content, primarily consider the proportion of content words, particularly adjectives and adverbs.

Development of a Prototype for an AI-Based Feedback Tool

The intended purposes of the AI-based feedback tool were threefold: (a) to assist teachers in providing feedback on students' writing, (b) to help students better prepare for the TOEFL Junior Writing test, and (c) to support the development of students' English writing skills in general. The prototype tool was primarily designed to assist teachers in providing feedback and helping students understand the feedback, although students also had access to view the feedback.

Feedback Dimensions

Overall, we attempted to provide feedback in the key dimensions of writing, as embedded in the rubrics: content, organization, and language use. Second language (L2) writing research has demonstrated that these areas are significant determinants contributing to high-quality writing (e.g., Biber & Gray, 2013; Crossley & McNamara, 2012; Cumming et al., 2005; Cushing Weigle, 2002; Kormos, 2011; Wolf et al., 2018). In addition, we included the aspects of mechanics (i.e., capitalization, punctuation, and spelling) as an added dimension in our feedback areas as they are important to learn for academic writing in school settings. Finally, we were cognizant of the amount of feedback provided to avoid overburdening students with excessive cognitive load. For instance, we identified the common grammatical error types from the previous TOEFL Junior Comprehensive test data and designed the tool to provide feedback in selected areas. Specifically, these areas included subject-verb agreement, verb forms, the

consistency between antecedents and their pronouns, prepositions, articles, and comparative word forms. The feedback presentation in these areas was designed for students to view one area at a time, using a togglable annotation function.

Our underlying approach to providing feedback was to closely link the feedback language to the holistic rubrics of the TOEFL Junior Writing tasks in an analytic manner. The intent of this approach was to inform teachers about the strengths and areas of improvement in students' writing as assessed against the TOEFL Junior Writing rubrics. Thus, our feedback content and language were formulated to correspond to the descriptors of the score points 1 to 4 on the rubrics regarding the content, cohesion and coherence, and language use dimensions. We also formulated the feedback language for the dimension of mechanics. As an example, the descriptors of each score point for the Opinion task rubric are presented in Appendix B. Taking the language use dimension for example, lexical variation is explicitly mentioned to differentiate between score levels (e.g., some, little, limited variation). Aligning these dimensions with their variations, we developed a list of two to four descriptors for each dimension as feedback describing the characteristics of the student's writing. These descriptors identified either the presence of specific features (e.g., position statement, cohesive elements) or the degree of linguistic diversity and content details (e.g., little, some, or wide range of diversity, some or many details). Table 3 displays a summary of the feedback aspects in each dimension for the Opinion task.

Table 3. A Summary of Feedback Aspects in Each Dimension of the Rubric (Opinion Task)

Writing dimension	Feedback aspect	Feedback characteristics
Content	Fluency	Words: 120 words (Expected length: 100–150 words)
	Stating a position	States a position
	Supporting details	Provides many supporting details
Cohesion and coherence	Expressions for time/sequence, cause/effect, contrast	Includes expressions showing time and sequencing, cause and effect, and contrast
	Pronouns	Includes no pronouns
	Linking words	Includes linking words that connect ideas (linking words can be highlighted in students' writing)
Language use	Lexical diversity	Includes a wide range of vocabulary for the task
	Syntactic diversity	Includes a narrow range of grammatical forms for the task
	Lexico-grammatical accuracy	Includes some errors at the word and phrase level For example: subject-verb agreement, prepositions
	Sentence structure	Includes some errors with sentence structure, such as run-ons or fragments
Mechanics	Spelling	90% of words spelled correctly
	Capitalization	Includes mostly accurate capitalization
	Punctuation	Does not include any final punctuation

Selection of AWE Features

In order to identify specific AWE engine features to generate feedback and determine cut points for the degree of specific feedback features, we utilized the operational data obtained from the TOEFL Junior Comprehensive test prior to 2017. We analyzed a sample of 3,534 responses to the same prompts included in the feedback prototype tool (1,317 for the E-mail task, 1,283 for the Opinion task, and 934 for the Listen-Write task). These prompts were retired from the TOEFL Junior Comprehensive test and used as TOEFL Junior Writing practice tasks for this prototype tool. We computed cross-tabulated descriptive statistics of microfeature values from response writings against their task scores. Preliminary analysis revealed that certain microfeature values (i.e., features used to compute other composite features) were not dispersed across the full range of task scores, whereas the normalized sums of conceptually related microfeatures were. Hence, relevant microfeatures were summed and normalized to be included as feedback features. For instance, the lexico-grammatical accuracy

feedback feature is the normalized sum of microfeatures that are counts of lexico-grammatical error types, such as subject-verb agreement errors. We also performed correlational analyses between the feature values and task scores to evaluate the significance of the features for writing quality. In sum, a sample of AWE features were selected based on the following criteria:

- Key writing dimensions specific to the TOEFL Junior Writing test: The features are conceptually aligned with the key writing dimensions and descriptors in the TOEFL Junior Writing scoring rubrics.
- Discrimination of writing quality: The feature values are dispersed across the full range of task scores, differentiating among writing qualities.
- Comprehensibility: The features are comprehensible to teachers and students for writing instruction.

Establishment of Cut Point Values

Once the features were selected, the range of feature values across the task scores were closely examined to determine the cut values to differentiate the degrees of feedback features. Initially, four levels of feedback descriptors were set to correspond to the TOEFL Junior Writing task rubric scores from 1 to 4. The median feature values for Score Groups 2, 3, and 4 were used as cut-off values. In other words, the cut values were selected to represent certain points on the scale that included the relevant descriptors and were used to determine which level of feedback descriptor was shown to users. In instances where the shape of the feature value distribution did not facilitate discriminating between task score points, two or three descriptors were used instead of four levels of descriptors. For features that represented negative qualities of writing (i.e., the normalized number of grammatical errors), the rules for determining the feedback level were reversed.

Table 4 illustrates the initial process of setting cut values based on feature values. For instance, for Feedback Level 4 in the Opinion task, the feature representing syntactic diversity has a cut value of 2.7. When the feature value (f in Table 4) associated with syntactic complexity is greater than or equal to 2.7 (corresponding to the median of feature value for task responses with a rubric score of 4), the descriptor linked to Feedback Level 4 is included on

the feedback report. On the other hand, for Feedback Level 1 in the Opinion task, the feature representing lexico-grammatical accuracy has a cut value of 6.0. Because this feature is based on negative writing qualities (i.e., counts of errors), feature values greater than 6.0 are linked to the Feedback Level 1 descriptor.

Table 4. Process for Setting Cut Points

Feedback level	Feature value (f) cut-off rules	
	Positive writing qualities	Negative writing qualities
1	f < rubric score 2 median of f	f > rubric score 2 median of f
2	f < rubric score 3 median of f	f > rubric score 3 median of f
3	f < rubric score 4 median of f	f > rubric score 4 median of f
4	f ≥ rubric score 4 median of f	f ≤ rubric score 4 median of f

While two to three levels of descriptors were provided for each feedback area, the numerical value of the feedback level (i.e., Feedback Level 1, 2, etc.) was not displayed to users. A first draft of feedback descriptors was created by borrowing from the language of the rubric descriptors for each task. Descriptor language was edited in cases where the number of feedback descriptors was reduced to less than four. Additional sets of descriptors, corresponding to mechanics features, were drafted in the same style. In some instances, multiple feedback descriptors were drafted to address a single rubric descriptor if there was not a feature that covered the entire writing trait described by the rubric descriptor or if the rubric descriptor described multiple traits.

In order to evaluate and revise the cut values and feedback descriptor language, we applied the established cut values and their associated feedback descriptors to a sample of the operational data used in this study. A team of three ETS staff (one assessment developer and two researchers) reviewed each set of responses grouped by task and task score to decide if the cut value should be adjusted to align with the feedback descriptor for the corresponding rubric score. Specific questions discussed during the review were regarding whether the feedback descriptor was appropriate for the sample writings and whether the cut values needed to be higher or lower to accurately reflect the sample writings. During this phase, feedback descriptor language was edited to more closely reflect the writing traits represented by the features and to be more interpretable by teachers and students in the intended context of the tool.

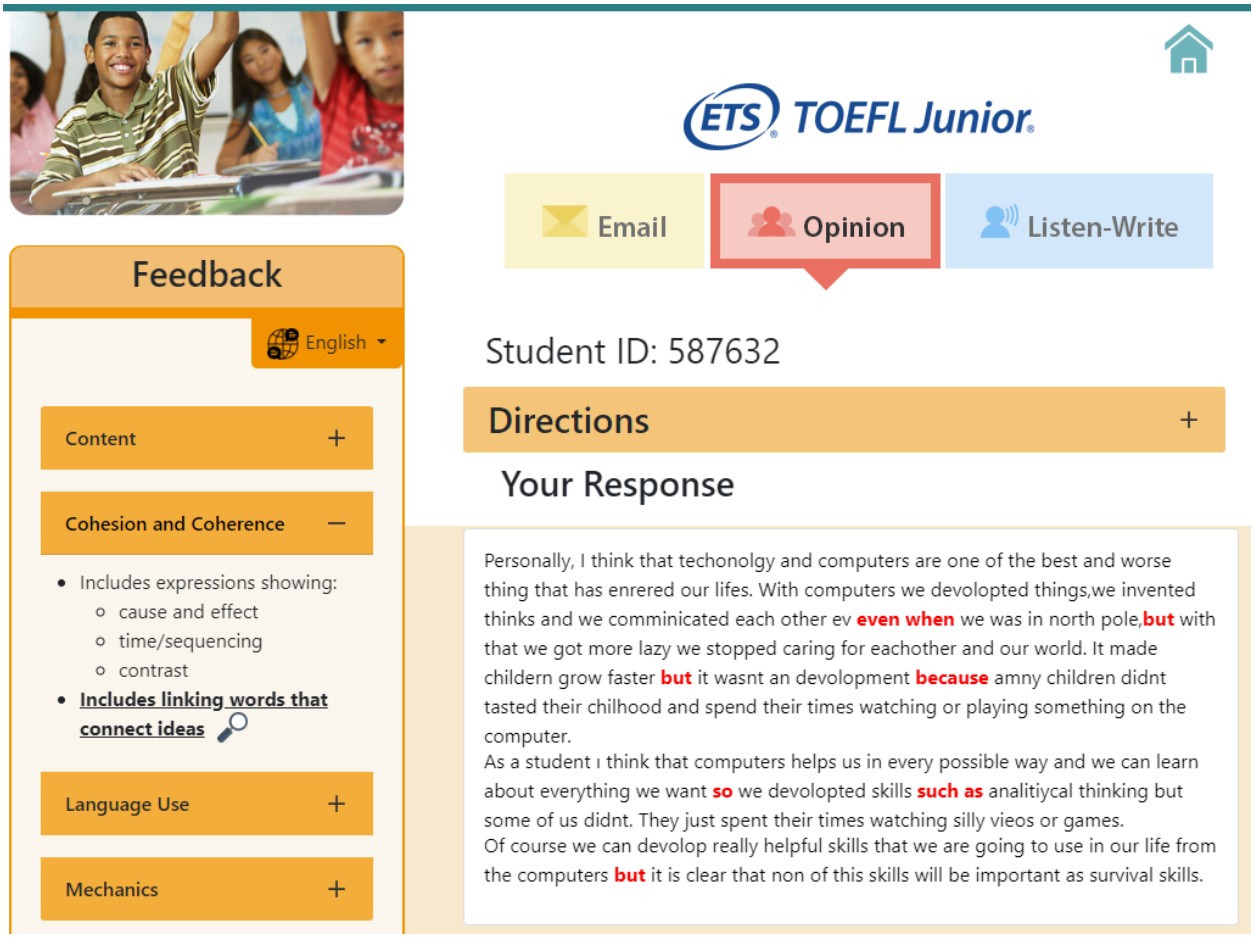
Additionally, it was decided to omit certain sets of descriptors from the E-mail and Listen-Write tasks that were thought to be less useful for those tasks. For instance, the descriptor addressing sentence structure was not included in the E-mail task feedback because using fragments is sometimes appropriate in this register. Similarly, the Listen-Write task did not include the descriptor for stating a position, as this was not required for the task.

To provide more granular feedback, additional information about the characteristics of responses is provided on the feedback report (e.g., togglable annotation of linking expressions used in the response, lists of error types in the response with togglable annotation, number of words in the response, number of spelling errors in the response). In particular, togglable annotations were made available for feedback on linking expressions and lexico-grammatical error types (see Figure 1). Based on the ETS corpus of linking expressions, the AWE engine was utilized to highlight the linking expressions in students' writings. Similarly, common error types such as subject-verb disagreement and inadequate prepositions were highlighted using the natural language processing capabilities of the AWE engine. The usefulness and interpretability of this information were reviewed by the project team while reviewing the descriptors to decide if it should be included in the feedback tool.

Structure of the Web-Based Prototype Tool

The web-based prototype tool consisted of two main components from the user's perspective: (a) TOEFL Junior Writing practice activities including three tasks and (b) a feedback report showing the writing prompts, students' responses, and automated feedback. The interface was designed to present information in smaller, more manageable chunks (e.g., elements that expand when the heading is selected). The feedback was available in different languages. For the usability study, we included three languages (English, Korean, and Turkish). Figure 1 presents a screenshot of the feedback report component to illustrate the design interface. This example shows feedback on the Coherence and Cohesion dimension for a response to the Opinion task with linking words highlighting toggled on.

Figure 1. Screenshot of the Feedback Report in the Prototype Tool With Linking Words in Boldface



Usability Study

To gather users' feedback about the prototype tool, we conducted a small-scale usability study. Specifically, we addressed the following research questions:

1. How do teachers perceive the usefulness of the AI-based feedback tool for students' writing skills?
2. How do students perceive the usefulness of the AI-based feedback tool to improve their English writing skills?
3. What suggestions do teachers and students have for further improvement of the tool?

Method

Participants

We chose South Korea and Türkiye for participant recruitment because the TOEFL Junior Comprehensive test was administered in these countries in high volumes. We first recruited five teachers from a private middle school in Türkiye and two teachers from after-school language institutes in South Korea on a voluntary basis. Teachers were asked to recruit two students from their classes within the target age range of 11–16 who had the ability to type a paragraph or more in English on a computer with the specification that the students should have different levels of English proficiency if possible. This approach resulted in a total of 14 students for the study. Written consent of the teachers and students' parents or guardians was collected prior to data collection; students were asked for their assent before participating in the interviews.

The recruited students were in middle-school grades and within the age range of 12–13, with the average age of 12.6. In general, the students from Türkiye had high English oral proficiency, as their school offered some courses in English-medium instruction in addition to intensive English courses. The students from Korea attended public schools where instruction was in Korean. The students also attended after-school language institutes, mainly for test preparation purposes. According to teachers' judgment, students from Türkiye had intermediate to high English writing proficiency, and students from Korea were at the intermediate English writing proficiency level. Based on the background questionnaire responses, 10 students (71%) indicated that they had previously taken the TOEFL Junior tests and seven (50%) indicated that they had previously taken the TOEFL Primary® test. The students also reported that they had been learning English for 7 years on average. Background information about individual students is included in Table 5.

Table 5. Student Background Information

ID	Age	Country	First	TOEFL tests previously taken	
				TOEFL Primary	TOEFL Junior
KS01	13	South Korea	Korean	No	No
KS02	13	South Korea	Korean	Yes	No
KS03	12	South Korea	Korean	Yes	No
KS04	12	South Korea	Korean	Yes	No
TS01	13	Türkiye	Turkish	No	Yes
TS02	13	Türkiye	Turkish	Yes	No
TS03	13	Türkiye	Turkish	No	No
TS04	13	Türkiye	Turkish	No	No
TS05	13	Türkiye	Turkish	Yes	Yes
TS06	13	Türkiye	Turkish	Yes	Yes
TS07	12	Türkiye	Turkish	Yes	Yes
TS08	12	Türkiye	Turkish	Yes	Yes
TS09	12	Türkiye	Turkish	Yes	Yes
TS10	13	Türkiye	Turkish	Yes	Yes

The teachers who participated in the study were English teachers from the selected institutions, with between 4 and 21 years of teaching experience (mean = 16.0). Four of the teachers had degrees related to English language teaching, while three had studied other disciplines. One teacher had a doctoral degree, two had master's degrees, and four had bachelor's degrees. The teachers from Türkiye conducted their instruction in English whereas the teachers from Korea used Korean in their instruction. Teacher background information is shown in Table 6.

Teachers were surveyed on the instructional practices they used in teaching English writing. The results indicate that all teachers required their students to use a computer or tablet for writing instruction at least once a month, with six teachers requiring it once a week. All participating teachers reported that students were typically asked to write at the paragraph level or to compose some manner of text (e.g., essays) in various genres (e.g., summarizing, narrative, argumentative).

Table 6. Teacher Background Information

ID	Country	Years teaching English	Highest degree earned	Degree area
KT01	South Korea	4	Bachelor's	English education
KT02	South Korea	20	Doctoral	Chemistry
TT01	Türkiye	17	Master's	English language and literature
TT02	Türkiye	21	Master's	Education
TT03	Türkiye	16	Bachelor's	Art history
TT04	Türkiye	19	Bachelor's	American culture and literature
TT05	Türkiye	15	Bachelor's	English

Study Instruments

In addition to the feedback prototype tool, a user manual for teachers was developed to provide guidance on how to use the tool with students. Surveys and interview protocols were developed for students and teachers to gather their background information and feedback about the prototype tool. The teacher survey and interview protocol included questions about background information, instructional practices, perceptions of the writing task, and perceptions of the feedback tool. The student survey and interview protocols included questions about background information, experiences with computerized writing, perceptions of the writing tasks, and perceptions of the feedback tool. With respect to the feedback tool, questions were focused on the degree of the tool's usefulness, the understandability of the feedback provided, the areas of useful feedback, and potential areas for improvement in the tool.

Procedure and Analysis

Teachers were given a website link to access the prototype tool, along with the user manual. They received individual logins for the tool, and the research team confirmed that their responses had been received after the trial. Any issues were resolved before proceeding with the administration to students. Teachers were explicitly informed that the feedback was generated using AI capabilities and that its accuracy was imperfect. Subsequently, teachers coordinated the administration of the TOEFL Junior Writing practice activity for the students from their school using the school's computer equipment at a time convenient for them. In a

single session, lasting approximately 1 hour, students completed the three writing tasks (i.e., E-mail, Opinion, and Listen-Write) and then convened with their teacher to review the feedback report (i.e., automated feedback) on their writing. After meeting with their teachers for feedback, students completed the online survey. Teachers completed the online survey at a later time after implementing the tool to their students.

After all students in each country completed the writing activity and teacher conference to review feedback, a series of group interview sessions were conducted virtually with (a) each teacher and their students and (b) all teachers from the country. Teacher interviews were conducted in English whereas student interviews were conducted in a mix of English and students' native languages with assistance from the teachers. Based on the interview protocols, a pair of researchers conducted each interview session, lasting 45 to 60 minutes. During the interviews, participants were shown their feedback reports to help elicit their opinions. Audio and video of the interviews were recorded for notetaking purposes. The interviews were also transcribed.

Descriptive statistics and response frequencies were computed for the student and teacher surveys. A pair of researchers reviewed the interview data and organized them according to protocol questions to summarize recurring themes and differences in users' feedback on the tool.

Findings

RQ1: Teachers' Perceptions

On the survey, all but one teacher reported that they found the feedback provided by the AI-based feedback tool to be useful for understanding students' strengths and weaknesses as well as for instructional planning, with the average response falling between "useful" and "somewhat useful" for all three questions (Table 7). One teacher reported that it was "not that useful" for both understanding students' strengths and instructional planning.

Table 7. Descriptive Statistics and Frequencies on Feedback Usefulness (*n* = 7 Teachers)

Purpose	*Mean*	*SD*	Very useful	Useful	Somewhat useful	Not that useful	Not useful at all
For understanding students' strengths	3.71	0.95	1	4	1	1	0
For understanding students' weaknesses	3.57	0.53	0	4	3	0	0
For instructional planning	3.29	0.76	0	3	3	1	0

Note. Very useful = 5; useful = 4; somewhat useful = 3; not that useful = 2; not useful at all = 1.

Similarly, perceptions on the usefulness of each type of feedback were generally positive, with average responses nearing or being exactly at "useful" for all types of feedback (Table 8). One teacher (TT04) found every type of feedback to be very useful. She said, "I would definitely use this system as a teacher . . . I totally like the system you created." When prompted to elaborate on her survey responses, she explained that the feedback descriptors—particularly those addressing cohesion and coherence—addressed objectives covered in the classes she teaches and that she "was surprised that . . . students could actually transfer that into their writing skills." Two of the teachers who found varying degrees of usefulness across feedback types expressed favorable opinions of the highlightable feedback showing the location of errors and linking words. Another teacher added that the feedback would be useful for reporting students' writing performance to their parents.

The same teacher who previously reported that feedback was not that useful for understanding students' strengths or instructional planning indicated that feedback on content and feedback showing where errors are in the response were not that useful either. During the interview, he explained that he would prefer if the feedback report included predictions of test scores and suggestions on how students could improve writing (e.g., vocabulary recommendations).

Table 8. Descriptive Statistics and Frequencies on the Usefulness of Feedback Types (*n* = 7 Teachers)

Feedback type	*Mean*	*SD*	Very useful	Useful	Somewhat useful	Not that useful	Not useful at all
Feedback on content	3.71	1.11	2	2	2	1	0
Feedback on cohesion and coherence	4.00	0.82	2	3	2	0	0
Feedback on language use	3.71	0.76	1	3	3	0	0
Feedback on mechanics	3.86	0.90	2	2	3	0	0
Feedback showing where errors are in the response	3.71	1.11	2	2	2	1	0
Feedback showing where linking words are in the response	4.00	0.82	2	3	2	0	0

Note. Very useful = 5; useful = 4; somewhat useful = 3; not that useful = 2; not useful at all = 1.

With respect to the understandability of feedback, all but one teacher indicated that they "mostly understood" or "completely understood" the information on the feedback report, with the mean response for each dimension lying between the two options (Table 9). Notwithstanding, three of the teachers who rated their understanding of the feedback favorably expressed doubt that their students would be able to understand all of the language used on the report, primarily in regard to the terms "cohesion and coherence" and "mechanics" as well as some of the metalanguage used in the language use descriptors, noting that they do not use these expressions during classroom instruction. However, one teacher believed that the highlightable feedback would help students understand the meaning of the language in the descriptors.

The same teacher who did not perceive the feedback to be entirely useful was the only respondent to select "understood somewhat" and "had difficulty understanding" for each question addressing the understanding of feedback. Contrary to his survey responses, he stated

in the interview that he would expect middle school students to understand the language on the feedback report.

Table 9. Descriptive Statistics and Frequencies on Understandability of Feedback (*n* = 7 Teachers)

Feedback dimension	*Mean*	*SD*	Completely understood	Mostly understood	Understood somewhat	Had difficulty understanding	Could not understand at all
Content	4.43	0.79	4	2	1	0	0
Cohesion and coherence	4.14	1.07	3	3	0	1	0
Language use	4.29	0.76	3	3	1	0	0
Mechanics	4.14	1.07	3	3	0	1	0

Note. Completely understood = 5; mostly understood = 4; understood somewhat = 3; had difficulty understanding = 2; could not understand at all = 1.

RQ2: Students' Perceptions

All students found the feedback tool to have some degree of usefulness for understanding their strengths and weaknesses in writing, with half finding it very useful for understanding their strengths (Table 10). The mean response to the two questions related to understanding strengths and weaknesses was between "useful" and "very useful," which suggests that the students' perceptions of the feedback tool were generally more positive than the teachers' perceptions. For example, one student (TS08) stated, "It shows what I should improve and what I've done good . . . Maybe I could add a bit longer and add more linking words."

Table 10. Descriptive Statistics and Frequencies on the Usefulness of Feedback (*n* = 14 Students)

Purpose	*Mean*	*SD*	Very useful	Useful	Somewhat useful	Not that useful	Not useful at all
For understanding what you are good at	4.36	0.74	7	5	2	0	0
For understanding what you need help with	4.21	0.80	6	5	3	0	0

Note. Very useful = 5; useful = 4; somewhat useful = 3; not that useful = 2; not useful at all = 1.

Concerning the types of feedback, half or more of the students found every type of feedback provided on the report to be very useful, on average rating each one between "very useful" and "useful" (Table 11). In the interview, when prompted to decide which dimension of feedback was most useful, the students generally chose cohesion and coherence or language use, mentioning that both categories provided information that was useful for improving their writing. In terms of cohesion and coherence, one student explained that being able to highlight the linking words used in her response helped her understand if she was overusing certain expressions. Another student shared that she did not realize that she had not used pronouns in her response until receiving feedback pointing this out. With respect to language use, two students mentioned that they appreciated being able to highlight language use errors, explaining why they favored this category. Another student noted that the descriptor addressing lexical range helped her understand that she should try to use a more diverse vocabulary.

However, not all students agreed that the language use, mechanics, and error-highlighting feedback were useful. One student thought that both language use and mechanics feedback were not that useful, elaborating in the interview that not all types of errors were included in these dimensions. Another student rated the error highlighting feature as not that useful. Further examination of this student's responses revealed that although he did make a few errors with spelling and sentence structure, unlike all but one of the other students, no

errors that were highlightable by the tool were made in any of his responses. Furthermore, the tool misclassified a single word as having a capitalization error and highlighted it, which the student noted in both the interview and the survey.

Table 11. Descriptive Statistics and Frequencies on Usefulness of Feedback Types (*n* = 14 Students)

Feedback type	*Mean*	*SD*	Very useful	Useful	Somewhat useful	Not that useful	Not useful at all
Content	4.36	0.74	7	5	2	0	0
Cohesion and coherence	4.50	0.76	9	3	2	0	0
Language use	4.43	0.94	9	3	1	1	0
Mechanics	4.29	0.91	7	5	1	1	0
Error highlighting	4.14	1.03	7	3	3	1	0
Linking word highlighting	4.21	0.89	7	3	4	0	0

Note. Very useful = 5; useful = 4; somewhat useful = 3; not that useful = 2; not useful at all = 1.

In regard to the understandability of feedback, every student reported that they understood, to some degree, all feedback types and the feedback overall (Table 12). For each feedback type and overall, at least half of all students reported that they completely understood. Only for feedback on cohesion and coherence did a student indicate that she understood a little.

During the interview, some students expressed difficulties in understanding the feedback information at first. Two students shared that they did not understand what "cohesion and coherence" meant when they first viewed the report but added that they understood the concept after viewing their feedback. When prompted to define cohesion and coherence, the one student (TS04) explained that it relates to linking words and how "smooth" the connection between paragraphs is, and the other (TS10) added that cohesion and coherence are related to the "the connect[ion] between sentences." A student in a separate

interview session added that she used the translation feature to learn the meaning of cohesion and coherence.

Some students also shared that they did not completely understand all of the language use descriptors. Five students mentioned that they did not understand the meaning of certain metalanguage used in the report, particularly "determiners," "run-ons," and "fragments," because they had not learned these terms in school, but stated that they understood other metalanguage that was covered in instruction, such as "articles" and "prepositions." The students mentioned that they toggled between English and their first language (Korean or Turkish) to understand those terms. When prompted, one student was able to provide a definition of "subject-verb agreement," and another (TS09) was able to explain that the descriptor stating that her response included a "narrow range of vocabulary" implied that she "use[s] words over and over again."

Table 12. Descriptive Statistics and Frequencies on the Understandability of Feedback (*n* = 14 Students)

Feedback area	*Mean*	*SD*	Completely understood	Mostly understood	Understood a little	Had difficulty understanding	Could not understand at all
Content	4.71	0.89	10	4	0	0	0
Cohesion and coherence	4.50	0.65	8	5	1	0	0
Language use	4.64	0.50	9	5	0	0	0
Mechanics	4.50	0.52	7	7	0	0	0
Overall	4.64	0.50	9	5	0	0	0

Note. Completely understood = 5; mostly understood = 4; understood somewhat = 3; had difficulty understanding = 2; could not understand at all = 1.

RQ3: Suggestions for Further Improvement of the Tool

Students and teachers offered their ideas for further improvement of the feedback tool throughout the interview and in response to the final question on both the survey and interview protocol asking what other types of feedback would be useful. Table 13 lists seven themes from student and teacher suggestions, providing sample comments from individual participants for each theme. Two themes (error correction and vocabulary recommendation) reflect the sentiments from students and teachers that the feedback provided by the tool should be more explicit not only by indicating the presence of errors or the limitations in writing performance (e.g., using a limited range of vocabulary), but also by providing recommendations (e.g., corrections of errors, synonyms to use in place of common words). Students and teachers further suggested that the feedback could be improved by revising descriptors to be more detailed and by adding more descriptors. In line with their generally favorable perceptions of the highlighting features, students and teachers both agreed that highlighting options should be included with more of the descriptors, particularly with the spelling descriptor, which only listed the percentage of words spelled correctly. Again, both groups were interested in receiving information about proficiency levels and suggested that it would be useful to provide test score information on the feedback report. Similarly, one student suggested that it would be useful to have information about how to improve his writing with reference to the levels of the CEFR.

The final theme, simplification, reflects a sentiment among teachers that the language used on the feedback report is too complex for students and would therefore be more useful if simplified; however, no students expressed this sentiment. Although all participants were asked to suggest improvements to the feedback report, it should be noted that not all participants thought that changes were necessary. The one teacher (TT04) who did not make any suggestions commented, “I would definitely like to use the system as a teacher while evaluating my students’ writing . . . I really liked the system you created.”

Table 13. A Summary of Users' Suggestion Themes for Feedback Tool Improvement

Themes	Student suggestion	Teacher suggestion
Error correction	It shows my mistakes, but like if it shows the correct forms of them, it will be better. (KS02, TS01, TS05, TS09)	If the system could show the correct answer, for example with the punctuation or spelling mistake or a grammatical structural problem … it could be really helpful. (KT01, TT01, TT02)
Vocabulary recommendation	It would also be good to add like synonyms and like other words to like help with the narrow choice of words. (TS03, TS06, TS09, TS10)	The feedback should also show what other words students can use. (KT01; translated from Korean)
Increased detail	I think the feedbacks could be more detailed. (KS01, KS02, TS05)	I was expecting to see more criteria, especially in the content section. (KT01, KT02, TT02)
Additional highlighting	It says the percentage of the words that I had written correct [in terms of spelling], but it doesn't show which words was incorrect. That would be more useful. (TS06)	I agree with the kids when they said they wanted to see their spelling mistakes. (TT05)
Score prediction	I think it would be interesting if you gave it a level. People are curious about their scores. (KS01, KS02, TS10; translated from Korean)	It should give a score. (KT01, TT02; translated from Korean)
Next steps	I think it would be better if there is a part like how can I write in an upper level. Like if my writing is B2, how can I write in C1 or something like that? (TS10)	-
Simplification	-	They're not familiar with the terminology, but apart from that, if there were more kid friendly explanations, that would have been fantastic. (TT01, TT02, TT03, TT05)

Note. The IDs noted after each quote include the participant to whom it was attributed and any participant who expressed agreement or made a similar suggestion.

Discussion

This study explored the use of ETS's AWE engine to provide automated feedback for teachers and students who prepare for the TOEFL Junior Writing test and aim to improve English writing skills. In particular, we designed the feedback to be closely linked to the key writing dimensions and proficiency descriptors outlined in the TOEFL Junior Writing rubrics, including content, cohesion and coherence, language use, and mechanics. In developing specific feedback for each key writing dimension, we considered both the capabilities of the AWE engine and the appropriate amount of the feedback. A feedback prototype tool was created and tested by a small number of teachers and students from Korea and Türkiye.

The usability study generally garnered positive feedback about the tool except for one teacher who desired the tool to provide scores for students' writings. Overall, teachers expressed their wish to make the tool available for instructional use, underscoring the importance of providing feedback for students' writing skills. Although the prototype tool was designed primarily for teachers, they suggested that students should be the direct viewers of the feedback, considering middle-school students' capacity for independent learning. Thus, teachers recommended refining the feedback language for students to understand, even though teachers could explain the feedback to students. Similarly, all students found the feedback tool useful. During interviews with researchers, students were able to analyze their writings in comparison to the provided feedback. It appears that reviewing their individualized feedback itself stimulated students' metacognitive strategies, which are important to become self-regulated learners and improve their L2 writing skills (Lee & Mak, 2018).

The areas suggested by teachers and students for improving the prototype tool revealed common patterns. Both groups strongly advocated for the tool to provide direct and corrective feedback. They valued the highlighting feature (i.e., togglable annotations) in the tool, as it pinpointed specific areas requiring attention. Because students were still learning various syntactic features and vocabulary in English, they favored corrective feedback for language use and mechanics as well as suggestions for expanding their vocabulary. It was interesting to note that students sought more feedback on content in addition to language use and mechanics.

Seemingly, students were enthusiastic about improving their English writing skills by seeking detailed feedback.

It is important to acknowledge the limitations of the study. First, the scope of this study was limited to exploring the capabilities of ETS's AWE engine in providing AI-based feedback for students' TOEFL Junior Writing practice tasks. The usefulness of the feedback tool was examined based on participants' perceptions, not by analyzing their revision processes in integrating the provided feedback. Secondly, the types of AI-based feedback were restricted by the capabilities of ETS's AWE engine. Thirdly, as an exploratory, prototyping study, the study included only a small group of students with intermediate and advanced English proficiency and their teachers. Hence, the findings should be interpreted with caution and limited to the context of this study.

Despite the limitations, the study provided valuable empirical evidence of the promising potential of utilizing the AWE engine to offer individualized feedback on students' writings for the TOEFL Junior Writing test. We hope that the prototype tool and suggestions provided by the study participants can serve as a useful foundation for further development and refinement of AI-based tools to support TOEFL Junior Writing test users. To enhance students' learning opportunities, it will be important for test providers to offer practice materials (in this study, a feedback tool) that help students improve their writing skills while they prepare for the test.

The findings and limitations described in this study point to some future research areas to refine the current prototype tool or to leverage AI capabilities to enhance the score report features of the TOEFL Junior Writing test. In terms of immediate research areas, it will be valuable to examine how students use the feedback provided from the tool to revise their writing. In this regard, the accuracy of the feedback will be an important aspect to be taken into consideration. Additionally, it will be crucial to investigate any additional types of feedback students and teachers desire from the tool. The present study explored this particular aspect with a small sample; however, future studies with larger samples from diverse backgrounds (e.g., various educational settings, a wide range of English proficiency levels) will provide valuable insights to further develop such AI-based feedback tools for learners and teachers.

References

Biber, D., & Gray, B. (2013). *Discourse characteristics of writing and speaking task types on the TOEFL iBT® test: A lexico-grammatical analysis* (TOEFL iBT® Research Report No. 19). ETS. https://doi.org/10.1002/j.2333-8504.2013.tb02311.x

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*(2), 115–135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, *10*(1), 5–43. https://doi.org/10.1016/j.asw.2005.02.001

Cushing Weigle, S. (2002). *Assessing writing.* Cambridge University Press.

ETS. (2019). TOEFL Junior® framework and test development. *TOEFL® Research Insight: Vol. 7.* https://ets.org/pdfs/toefl/toefl-ibt-insight-s1v7.pdf

Fu, Q.-K., Zou, D., Xie, H., & Cheng, G. (2024). A review of AWE feedback: Types, learning outcomes, and implications. *Computer Assisted Language Learning*, *37*(1–2), 179–221. https://doi.org/10.1080/09588221.2022.2033787

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1) 81–112. http://www.jstor.org/stable/4624888

Hoang, G. T. L., & Kunnan, A. J. (2016). Automated essay evaluation for English language learners: A case study of MY Access. *Language Assessment Quarterly*, *13*(4), 359–376. https://doi.org/10.1080/15434303.2016.1230121

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing*, *20*(2), 148–161. https://doi.org/10.1016/j.jslw.2011.02.001

Lee, I., & Mak, P. (2018). Metacognition and metacognitive instruction in second language writing classrooms. *TESOL Quarterly*, *52*(4), 1085–1097. https://www.jstor.org/stable/44987051

Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and

usefulness of feedback as part of argument-based validation. *Educational Psychology*, *37*(1), 8–25. https://doi.org/10.1080/01443410.2015.1136407

So, Y., Wolf, M. K., Hauck, M. C., Mollaun, P., Rybinski, P., Tumposky, D., & Wang, L. (2015). *TOEFL Junior® design framework* (TOEFL Junior® Research Report No. 2). ETS. https://doi.org/10.1002/ets2.12058

Wilson, J., Huang, Y., Palermo, C., Beard, G., & MacArthur, C. A. (2021). Automated feedback and automated scoring in the elementary grades: Usage, attitudes, and associations with writing outcomes in a districtwide implementation of MI Write. *International Journal of Artificial Intelligence in Education*, *31*, 234–276. https://doi.org/10.1007/s40593-020-00236-w

Wolf, M. K., Oh, S., Wang, Y., & Tsutagawa, F. S. (2018). Young adolescent EFL students' writing skill development: Insights from assessment data. *Language Assessment Quarterly*, *15*(4), 311–329. https://doi.org/10.1080/15434303.2018.1531868

Appendix A. The Descriptor of the Highest Score Point of Each Task Rubric

E-mail Task

A typical response at this level is characterized by the following:

- responds to all questions in the e-mail, directly or indirectly
- is coherent
- shows lexical variation appropriate for the task
- displays a varied sentence structure appropriate for the task
- may contain minor errors but they do not interfere with meaning

Opinion Task

A typical response at this level is characterized by the following:

- states a position on the topic
- provides support for the position, with specific details and/or examples
- is mostly well organized and coherent
- shows lexical variation appropriate for the task
- displays a varied sentence structure appropriate for the task
- may contain minor errors but they do not interfere with meaning or clarity

Listen-Write Task

A typical response at this level is characterized by the following:

- accurately provides all key points
- provides support using relevant details from the talk
- is mostly well organized and coherent
- shows lexical variation appropriate for the task
- displays a varied sentence structure appropriate for the task
- may contain errors but they do not interfere with meaning or clarity

Appendix B. TOEFL Junior Writing Scoring Guide for the Opinion Task

Score	Development and Language Use Descriptors
4	**A typical response at this level is characterized by the following:** • states a position on the topic • provides support for the position, with specific details and/or examples • is mostly well organized and coherent • shows lexical variation appropriate for the task • displays a varied sentence structure appropriate for the task • may contain minor errors but they do not interfere with meaning or clarity
3	**A typical response at this level is characterized by the following:** • states a position on the topic • provides support for the stated position, but may have difficulty doing so fully • is generally well organized, with an occasional lapse of clarity when connecting ideas • shows some lexical variation appropriate for the task • may display some variation in sentence structure appropriate for the task • may contain some errors that occasionally interfere with meaning
2	**A typical response at this level is characterized by the following:** • states a position on the topic, but provides inadequate/incomplete support, OR • only vaguely implies a position on the topic, and provides inadequate/incomplete support • connections between ideas are attempted, but are sometimes unclear or missing • shows little lexical variation (e.g., vocabulary is simple and repetitive), or frequently uses vocabulary incorrectly • shows little variation in sentence structure (e.g., sentences are mostly simple and short), and shows little control of sentence structures • may contain errors that frequently interfere with meaning
1	**A typical response at this level is characterized by the following:** • states a position but provides incoherent or no support OR • does not state a position, or makes only a minimal connection to the prompt and provides minimal or no support • is generally unorganized and incoherent • displays extremely limited vocabulary that is frequently used incorrectly • uses mostly incorrect sentence structures • displays many errors that seriously interfere with meaning
0	**Only copies words from the prompt, rejects the prompt, is completely off topic, consists of keystroke characters, is written in a foreign language, or is blank.**