Korean Journal of English Language and Linguistics, **Vol 26, April 2026, pp. 580-605**
DOI: 10.15738/kjell.26..202604.580

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

Developing and Validating a Curriculum-Based Rating Scale for Korean Middle School EFL Writing: An Argument-Based Validation Approach

Haeyun Jin (Korea National Open University) · **Wooyeon Kim** · **Sun-Young Oh** (Seoul National University)

Received: March 16, 2026
Revised: April 9, 2026
Accepted: April 9, 2026

Jin, Haeyun (First author)
Assistant Professor, Department of English Language and Literature
Korea National Open University
Email: haeyunj@knou.ac.kr

Kim, Wooyeon
Doctoral Student, Department of English Language Education
Seoul National University
Email: hallokatze@snu.ac.kr

Oh, Sun-Young (Corresponding author)
Professor, Department of English Language Education & Learning Sciences Research Institute, College of Education
Seoul National University
1 Gwanak-ro, Gwanak-gu
Seoul, Korea
Tel: +82-2-880-7675
Email: sunoh@snu.ac.kr

* This work was supported by the Learning Sciences Research Institute at Seoul National University (0767-20240007).

ABSTRACT

Jin, Haeyun, Wooyeon Kim and Sun-Young Oh. 2026. Developing and validating a curriculum-based rating scale for Korean middle school EFL writing: An argument-based validation approach. *Korean Journal of English Language and Linguistics 26*, 580-605.

Advances in artificial intelligence have led to the rapid expansion of automated writing evaluation tools for second language writing. However, the effectiveness of such systems depends critically on the assessment frameworks used to generate training data and interpret scores. In particular, little research has examined the development and empirical functioning of curriculum-based rating scales designed for secondary-level learners. Situated within a larger project developing an AI-assisted writing feedback tool for middle school English learners, this study reports the development and validation of a rating scale for assessing middle school EFL writing. Guided by an argument-based validation framework (Knoch and Chapelle 2018), the study examines the domain definition and evaluation inferences relevant to rater-mediated assessment. The scale was developed based on analyses of the national curriculum, middle school textbooks, and learner writing data. Its functioning was examined using Many-Facet Rasch Measurement (MFRM). The results indicate that the rating scale reflects key dimensions of middle school EFL writing and that scale criteria functioned in a consistent manner across raters. Category analyses further showed that the five score levels distinguished meaningful differences in student performance. The findings provide empirical support for the rating scale and offer methodological insights for the development of rating frameworks for future AI-assisted writing assessment systems.

KEYWORDS

rating scale development, argument-based validation, EFL writing assessment, Many-Facet Rasch measurement, AI-assisted writing assessment

1. Introduction

Recent advances in artificial intelligence have accelerated the development of technology-based tools designed to support second language (L2) writing. Automated writing evaluation (AWE) systems, in particular, provide immediate and personalized feedback that supports independent revision and classroom instruction (Cotos 2014, Shermis and Burstein 2003). Commercial platforms such as *Criterion, Pigai*, and *Grammarly* typically offer instant feedback, multiple drafting opportunities, and performance tracking, expanding access to individualized writing support (Woodworth and Barkaoui 2020). Despite these developments, questions remain regarding the educational grounding of AI-based writing tools. It is often unclear whether such systems are systematically aligned with curriculum-based achievement standards or calibrated to learners' developmental stages (Link et al. 2022).

These concerns are particularly consequential for low-proficiency learners, who typically have fewer opportunities for sustained writing practice and structured feedback. Research indicates that low-proficiency learners may feel overwhelmed by the volume and complexity of automated feedback, particularly when it is not tailored to their developmental stage, potentially widening achievement gaps rather than reducing them (Aluthman 2016, Koltovskaia 2020, Stevenson and Phakiti 2019, Tian and Zhou 2020). Moreover, the pedagogical effectiveness of AI systems depends not only on algorithmic sophistication but also on the representativeness of the learner data and the assessment frameworks on which they are trained (Whang et al. 2023). Existing L2-oriented tools largely target university-level writers (e.g., *CyWrite*, Cotos 2014, *Pigai*, Huang and Renandya 2020), leaving a relative absence of systems grounded in secondary learners' actual writing performance. Taken together, these issues highlight the need for AI-based writing feedback systems that are explicitly aligned with curriculum standards and empirically grounded in secondary learners' writing data.

Addressing this gap requires a clearly articulated and empirically validated assessment framework. However, even when curriculum-based rating scales are developed, their validation is often insufficient. Many studies describe scale development processes without systematically examining how scales function when applied by multiple raters in real assessment contexts (Mendoza and Knoch 2018). A rating scale that appears conceptually sound does not automatically function as intended. While many scale development studies rely primarily on traditional inter-rater reliability indices, these statistics alone cannot determine whether raters differ systematically in severity or whether score categories meaningfully distinguish levels of performance. Without such evidence, the validity of score-based inferences remains uncertain. This issue is particularly consequential in AI-based writing evaluation, as automated systems trained on rating data inevitably reflect the strengths and weaknesses of the underlying human scoring process. Rigorous empirical validation of rating scales is therefore essential for responsible AI-supported assessment.

Situated within a larger project developing an AI-based writing assessment tool tailored to middle school L2 English learners, particularly those with limited proficiency, the present study reports the development and empirical validation of a curriculum-based rating scale. Drawing on an argument-based validation framework (Knoch and Chapelle 2018), the study documents the scale's development grounded in learner data and examines its operational functioning using Many-Facet Rasch Measurement (MFRM), with particular attention to rater consistency and category functioning. By establishing empirical evidence for the scale's validity, this study lays the groundwork for the subsequent development of AI-supported AWE tools based on empirically grounded assessment frameworks.

2. Theoretical Framework

2.1 Existing Approaches to Rating Scale Development

The development of rating scales for writing assessment is a complex and consequential process. Rating scales do not simply function as scoring tools; rather, they operationalize the *construct* of writing ability and mediate the relationship between theoretical definitions of proficiency and observable learner performance (McNamara 2002). Decisions made during scale design, therefore, shape how writing ability is conceptualized, what features of performance are attended to, and how differences between learners are represented. Given this central role, the sources and procedures underlying rating scale development require careful attention.

In the literature on language assessment, rating scale development has been conceptualized in different ways. In general descriptions of test development, rating scale design has often been portrayed as drawing primarily on expert judgment informed by research or existing frameworks (Kane et al. 2017). In this view, scale descriptors are typically formulated on the basis of theoretical considerations, prior models of proficiency, professional experience, and may subsequently be refined through empirical work. Such descriptions reflect a long-standing tendency in language assessment to treat rating scale development as largely guided by expert intuition.

Fulcher (1987, 2003, 2012) offered a more critical perspective by distinguishing between measurement-driven and performance-driven approaches to rating scale construction. In Fulcher's terms, *measurement-driven scales* originate primarily from expert intuition, adaptations of existing proficiency frameworks, or generalized "can do" statements that are not directly derived from systematic analyses of actual learner performance. By contrast, *performance-driven scales* are constructed on the basis of empirical analyses of authentic performance data, with descriptors grounded in observable features of learner language. Fulcher (1987, 2012) argued that scales developed without a firm empirical foundation risk weakening construct validity, particularly when descriptors are abstract, decontextualized, or insufficiently aligned with real-world language use. Empirical analysis of performance data, in this sense, is not merely supplementary but central to establishing meaningful interpretations of scores.

This measurement-versus-performance distinction has been highly influential, but subsequent work has suggested that actual scale development practices are more varied than a strict dichotomy would imply. Montee and Malone (2014), for example, identified multiple approaches to scale construction, including a priori criteria adapted from existing standards, empirically derived scales based on performance data, theory-based scales grounded in models of language proficiency, and scales derived from learning goals and curricula. Rather than advocating for a single method, they argued that combining approaches may help mitigate the limitations associated with relying exclusively on any one source. Similarly, Knoch et al. (2021), in their review of published studies, observed that rating scales frequently draw on both *test-external* sources (e.g., curriculum standards, theoretical frameworks, prior research, expert input) and *test-internal* sources (e.g., analyses of learner texts, rater feedback, task characteristics). In practice, many scales are developed through iterative processes that integrate these different forms of evidence.

The existing body of work suggests a shift in the field from viewing rating scale development as primarily intuition-based toward recognizing the importance of empirical grounding, while also acknowledging that scale design typically involves multiple sources of input. At the same time, scale development studies do not always specify how these sources were integrated or how development decisions were justified, making it difficult to evaluate the extent to which descriptors are systematically grounded in learner performance and aligned with the intended construct.

Building on these discussions, the present study adopts a mixed approach to rating scale development. The scale

was informed by test-external sources, including national curriculum achievement standards, instructional materials, relevant research, and expert input, as well as test-internal sources derived from systematic analyses of middle school learners' writing and rater feedback. By integrating these sources, the study seeks to ensure that the resulting descriptors are both aligned with curriculum-based expectations and grounded in observable features of learner performance within the target context.

2.2 Argument-Based Validation

The present study draws on an argument-based approach to validation as the conceptual framework for guiding the development and validation of the rating scale. Although rating scale development and validation have been widely discussed in language assessment research, validation procedures are often not explicitly articulated in scale development reports (Mendoza and Knoch 2018). Moreover, relatively few rating scale studies have been situated within an explicit theoretical model of validation, making it difficult to evaluate how different sources of evidence contribute to the validity of score-based interpretations (see also Janssen et al. 2015, Knoch 2009, Youn 2015).

In response to these concerns, this study adopts an argument-based view of validation, which conceptualizes validity as the evaluation of score interpretations and uses rather than as a property of a test or a rating scale itself (Kane 2013, Messick 1989). Within this perspective, validation involves examining the chain of inferences that connects observed performances to the conclusions drawn from assessment results, and empirically supporting the assumptions underlying each inference. Specifically, this study draws on the validation framework articulated by Knoch and Chapelle (2018), which extends argument-based validation to the context of rater-mediated performance assessment. Their framework explicitly incorporates rating processes into the validity argument by identifying key inferences related to scale design, rater behavior, and score interpretation, along with associated warrants, assumptions, and potential sources of backing. This framework provides a systematic basis for evaluating whether a rating scale functions as intended when applied by multiple raters to actual learner performances.

Following this framework, the current study represents a partial validation effort focusing on two central inferences that are particularly relevant to the development and application of a rating scale: the domain definition inference and the evaluation inference. For each inference, the study articulates the corresponding warrant and examines the plausibility of underlying assumptions using empirical evidence. Evidence for the domain definition inference is drawn from the scale development process, including document analysis and examination of learner writing samples, while evidence for the evaluation inference is obtained mainly through the Many-Facet Rasch Measurement (MFRM) analysis.

The specific inferences, warrants, and assumptions addressed in this study are summarized in Table 1. Guided by this framework, the study addresses the following research questions, each corresponding to a warrant, which collectively examine the representativeness of the rating scale and its functioning in actual rating contexts.

1. To what extent does the newly developed rating scale reflect key dimensions of middle school EFL writing?
2. To what extent do raters apply the rating scale consistently and without unacceptable severity differences?
3. How well do the rating scale categories function as intended in distinguishing different levels of students' writing performance?

Table 1. Inferences, Warrants, Assumptions, and Backing for the Present Study (Based on Knoch and Chapelle 2018)

Domain definition: The rating scale appropriately represents the scoring criteria and performance features emphasized in the national curriculum and manifested in middle school EFL writing tasks.		
Warrants	Assumptions	Backing from this study
1. The scoring criteria and descriptors are designed to align with key assessment dimensions of middle school EFL writing, as reflected in curricular standards, textbooks, and actual student performance.	1. Core assessment dimensions can be identified through analysis of the national English curriculum achievement standards and major assessment descriptors.	1. Review of curriculum documents specifying writing-related achievement standards and major assessment descriptors.
	2. The scoring criteria can be derived from document analysis of existing textbooks.	2. Textbook analysis to extract recurring assessment foci and criteria.
	3. Scale descriptors adequately capture performance features identified in representative student writing samples across proficiency levels.	3. Empirical sampling and qualitative feature extraction from student writings.
	4. Descriptor clarity and relevance can be affirmed through domain expert review.	4. Focus group interviews with in-service teachers and resulting revisions
Evaluation: Observations are evaluated using procedures that provide observed scores with intended characteristics.		
Warrants	Assumptions	Backing from this study
2. Observed ratings reflect consistent application of the rating scale by raters.	1. Raters can consistently apply the scale to test tasks.	1. Inter-rater reliability index (Krippendorff's alpha), MFRM analysis (rater fit statistics)
	2. Differences in rater severity are within acceptable limits.	2. MFRM analysis (severity logit range)
3. The rating scale functions as designed, adequately differentiating levels of writing proficiency.	1. The scale is able to spread examinees into different levels and shows an acceptable category structure.	1. MFRM: Rating scale category analysis

3. Methods

3.1 Phase 1: Rating Scale Development (Domain Definition Inference: RQ1)

3.1.1 Approach to rating scale development

The rating scale development was informed by a hybrid approach to scale development, which integrates test-external sources with test-internal evidence derived from actual student performance (Montee and Malone 2014). *Test-external* sources included national curriculum achievement standards, middle school English textbooks, prior research on L2 English writing assessment, and expert input from experienced teachers. *Test-internal* sources consisted of student writing samples and ratings obtained during pilot rating. Integrating these sources supported the development of a rating scale that reflects curriculum-based expectations while capturing salient features of students' writing performance.

3.1.2 Scale development procedures

The development of the rating scale proceeded through several iterative stages. Table 2 summarizes how various sources informed each stage. First, a document analysis was conducted to examine prior studies on young learners' English writing ability, writing assessment materials from 13 middle school English textbooks, and writing-related achievement standards in the 2022 Revised National English Curriculum of Korea (Korean Ministry of Education 2022). Prior research on young learners' writing was interpreted to include studies on young and pre-adolescent learners in EFL writing contexts, and the studies reviewed were selected from well-established journals in language assessment (e.g., *Language Assessment Quarterly*, *Language Testing, Assessing Writing*) as representative and frequently cited work on the assessment of children's or young learners' writing. These studies were used to inform general dimensions of early-stage L2 writing ability, which were then considered in relation to the target population and context of middle school EFL writing. The analysis aimed to identify broad assessment domains and recurring emphases in the evaluation of middle school English writing. Based on this review, preliminary scoring criteria were established at a conceptual level, reflecting key dimensions of writing ability targeted in the curriculum. These criteria formed the basis of the initial rating scale, which was subsequently refined through empirical analysis of student writing samples.

To ground the rating scale in actual student performance, an empirical analysis of student writing samples was conducted during scale development. Two members of the research team served as scale developers. A subset of writing samples was randomly selected, stratified by school and grade level to ensure representativeness. Twenty students were selected, and both Task 1 and Task 2 were included, yielding 40 writing samples. Samples judged to represent anomalous cases (e.g., responses written largely in Korean) were excluded and replaced to avoid anchoring descriptors to extreme cases. The sampled texts were first roughly ordered according to overall writing quality to establish an initial performance continuum. The two researchers then examined the texts independently and subsequently compared their judgments to identify salient textual features distinguishing adjacent performance levels and to refine descriptor wording. Through this negotiated review process, additional indicators were incorporated into the descriptors to better capture empirically observed differences in student performance. The refined descriptors were applied to the same set of pilot samples in a second round of review, after which the researchers discussed remaining discrepancies and made final adjustments. This process resulted in a draft rating scale for subsequent expert review.

Table 2. Summary of Evidence Sources and Their Roles in Rating Scale Development

Source	Evidence examined	Role in scale development	Reflected in the final scale
Previous studies and existing writing tests	Assessment emphases	Informed the identification of broad assessment domains	Initial criteria (Content / Language Use / Spelling)
Assessment tasks from English textbooks	Evaluation emphases	Supported alignment with classroom-based evaluation practices	
Curriculum achievement standards	Writing-related achievement standards	Supported alignment with curriculum-relevant dimensions	
Learner texts	Observed performance features across levels	Informed refinement of level descriptors	Refined level descriptors
Focus group review	Teacher/expert feedback	Supported adjustment of descriptor wording and usability	Finalized level descriptors

Finally, the draft rating scale was reviewed by three middle school English teachers and a researcher specializing in diagnostic feedback tool development through an online focus group discussion. The session involved a structured discussion of the draft descriptors, with feedback focusing on descriptor clarity, alignment with curriculum expectations, and practical usability in classroom assessment contexts. Based on this feedback, the rating scale was revised and finalized for use in the main rating phase.

3.2 Phase 2: Main Rating and Quantitative Validation (Evaluation Inference: RQ2, RQ3)

3.2.1 Participants

3.2.1.1 Raters

Five in-service English teachers participated as raters. The raters were recruited through the research team's professional networks, taking into account their areas of interest and professional backgrounds, and consisted of teachers who expressed willingness to participate in the study. All were teaching English at middle or high schools at the time of data collection and held bachelor's and master's degrees in English education. Their teaching experience ranged from 4 to 15 years. Four raters were female and one was male.

3.2.1.2 Examinees

Writing samples were collected from 335 middle school students (233 from School A and 102 from School B), attending two public schools located in Seoul. Students in School A represented Grades 1 through 3, whereas students in School B were all in Grade 3. The participating schools were selected through professional networks of English teachers involved in the project, taking into account school-level achievement, feasibility of participation, and institutional support. Data were collected during regular English class sessions with the cooperation of classroom teachers. Prior to data collection, students and their parents were informed about the study, and data from only those who provided consent were included in the analysis. Before producing the writing samples, students completed a brief background questionnaire (see Table 3).

Table 3. Summary of Examinees' Background Questionnaire Results

Survey item	Response summary
Length of residence in English-speaking countries (months)	Mean = 0.66, *SD* = 5.51, Min = 0, Max = 90, Median = 0
Years of English study (years)	Less than 5 years: 40.61%, 5–6 years: 26.36%, 7–8 years: 24.24%, 9 years or more: 8.79%
English study outside school (weekly hours)	None: 24.24%, 1–2 hours: 26.06%, 3–4 hours: 23.94%, 5+ hours: 25.76%
Experience writing English texts longer than 40 words	Never: 20.30%, Once or twice: 43.03%, Several times: 32.42%, Frequently: 4.25%

The questionnaire results provide a descriptive profile of the participants as Korean middle school learners. The median length of overseas residence was 0 months, showing that most students had no immersion experience. Around two-thirds had studied English for fewer than seven years, consistent with expectations for this age group. Students also reported considerable variation in extracurricular English study time, ranging from none to five hours

per week. Prior experience with extended English writing (40+ words) appeared to be fairly limited. Most students reported having done this only once or twice (43.03%) or several times (32.42%), while relatively few indicated frequent experience (4.25%), and 20.30% reported no prior experience. This demographic profile provides important context for interpreting the writing performances evaluated in the present study.

3.2.2 Materials

3.2.2.1 Writing tasks

Two writing tasks were developed for this study based on the 2022 Revised English Curriculum of Korea and assessment practices in Korean middle schools. Both tasks were designed as short personal-experience narratives, a genre considered accessible for early-stage L2 writers and aligned with curriculum achievement standards emphasizing familiar topics (Hudson and Shapiro 1991). Draft versions of the tasks were reviewed by in-service English teachers to ensure clarity and appropriateness for the target population.

The two tasks shared the same structure and requirements but differed in topic. Task 1 asked students to write about three things they like and explain their reasons, whereas Task 2 asked students to describe the best day in their life and explain why it was memorable. For both tasks, the prompt was provided in English with a Korean translation, and included a brief "Required information" checklist guiding students toward essential content (Figure 1). Students were instructed to produce at least 40 words, reflecting the amount of writing typically expected within a single class period in Korean middle schools.

Students completed both tasks on school-issued tablet PCs using a custom-built online writing platform developed for this study. The platform allowed teachers to distribute prompts via a class link and enabled students to compose and submit responses directly on their devices. To preserve the authenticity of students' writing, spell-checking, auto-correction, and suggestion functions were disabled, and only a word-count display was visible while writing. Each task was completed within a 20-minute time limit.

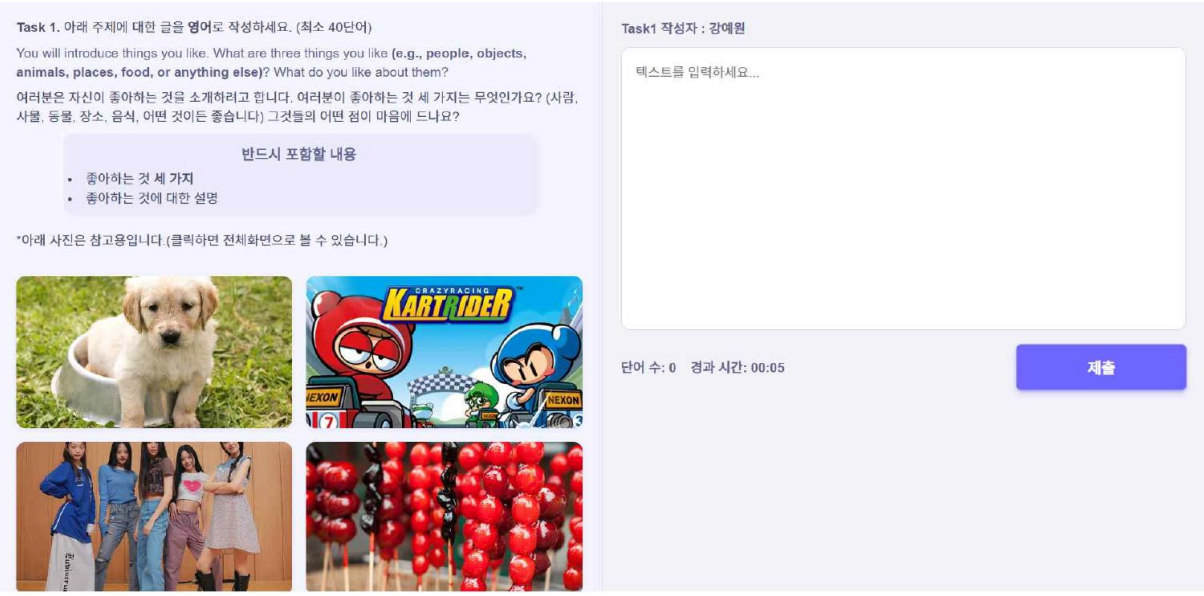

Figure 1. Screenshot of Task 1 on the Online Writing Platform

3.2.2.2 Writing samples

A total of 335 students from Schools A and B consented to participate in the study, and each student completed both writing tasks. The initial dataset consisted of 670 written texts. During preliminary screening, responses were excluded if they (a) contained no written content, (b) were written entirely in Korean, or (c) consisted of extremely short productions (e.g., one or two English words). After this initial screening, 324 and 320 responses were retained for Task 1 and Task 2, respectively. In addition, during the rating process, raters identified a number of responses as unscorable (e.g., "no content," "too short to evaluate"), and these responses were excluded from subsequent analyses. After applying both screening steps, the final dataset used for MFRM analysis consisted of responses from 315 examinees, each of whom had completed both tasks. A summary of the data screening process and the final analytic sample is presented in Table 4.

Table 4. Overview of Data Screening and Final Sample

Stage	Description	*N*
Initial dataset	335 students completed two tasks	670 texts (Task 1: 335, Task 2: 335)
Researcher screening	Invalid responses removed (no content, Korean-only, extremely short)	644 texts (Task 1: 324, Task 2: 320)
Rater screening	Unscorable responses excluded based on raters' judgments	630 texts (Task 1: 315, Task 2: 315)

3.2.3 Data collection procedures

Before scoring began, all raters participated in a 90-minute online training session to familiarize themselves with the rating scale and promote consistent scale application. All raters had prior experience with analytic writing assessments. The session included a review of the writing tasks and rating scale descriptors, a discussion of seven benchmark samples representing different performance levels, and an independent rating of five additional practice essays. A guided calibration discussion followed, allowing raters to align their interpretations of the scale and develop a shared understanding of the rating criteria.

After the rater training, each rater completed the scoring independently using a custom-built online rating platform developed for this study. As shown in Figure 2, the interface presented one student writing sample per screen. The interface displayed the student essay and analytic rating scale with selectable score categories (A–E) for each criterion. An additional option labeled "F" was provided to mark responses that were unscorable.

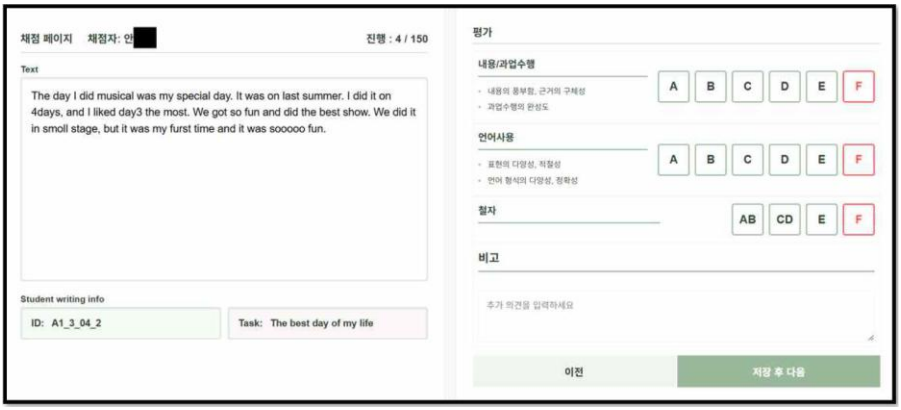

Figure 2. Screenshot of the Rating Platform

All writing samples were double-scored, with each essay evaluated independently by two raters. To minimize rater bias, writing samples were assigned in a counterbalanced manner so that each rater scored a comparable number of samples across schools and grade levels. The design ensured sufficient overlap among rater pairings to support a properly connected MFRM analysis (McNamara 1996). The order in which Task 1 and Task 2 essays appeared was varied across raters to control for potential order effects.

3.2.4 Data analysis

To address Research Questions 2 and 3, inter-rater reliability analysis and Many-Facet Rasch Measurement (MFRM) analysis were conducted. As an initial step for Research Question 2, overall inter-rater reliability was examined using Krippendorff's alpha (K-alpha). K-alpha was selected for several methodological reasons. First, the present study employed a partially crossed rating design (Eckes 2015), in which not all raters evaluated all writing samples. Under such conditions, traditional reliability indices (e.g., Cronbach's alpha) that assume fully crossed designs are not appropriate. K-alpha allows estimation of inter-rater agreement with partially overlapping data. In addition, K-alpha is well suited for ordinal rating data and can accommodate missing ratings, which are common in large-scale rating designs. K-alpha coefficients were computed using custom Python scripts.

Second, rater performance was examined using MFRM analysis conducted with FACETS (Version 4.1.8). Two models were estimated. Model 1 included all three criteria (Content, Language, and Spelling) to examine overall rating patterns and rater behavior across the full scale. Based on the preliminary MFRM findings, Model 2 excluded the Spelling criterion and focused on Content and Language, which demonstrated acceptable reliability. Model 2 served as the basis for the main analysis of rater severity and rater fit. Preliminary assumption checks, including examinations of unidimensionality and overall model fit, indicated no violations of assumptions for conducting the MFRM analyses.

For Research Question 3, the MFRM results from Model 2 were further analyzed to evaluate rating scale functioning. The analysis examined rating scale category statistics, examinee separation (strata), threshold ordering, and category probability curves to determine whether the score categories meaningfully distinguished levels of student writing performance in accordance with established guidelines for well-functioning rating scales (Eckes 2015).

4. Results

4.1 Representativeness of the Rating Scale Criteria and Descriptors (RQ1)

Research Question 1 examined the extent to which the newly developed rating scale reflects key dimensions of middle school EFL writing. This section reports key outcomes from the rating scale development that are relevant to the domain definition inference (RQ1).

4.1.1 Evidence from document analysis

As part of the rating scale development process, multiple sources of evidence were reviewed, including prior research on young learners' English writing assessment, commonly used standardized writing scales, middle school English textbooks, and the 2022 Revised National English Curriculum. The purpose of this analysis was to

identify broad assessment domains and recurring emphases in the evaluation of middle school English writing, which served as the conceptual foundation for the initial rating scale.

First, findings from prior research on young learners' English writing ability indicated that a limited set of dimensions has been consistently emphasized across grade levels and task types. Studies focusing on elementary and middle school learners frequently assessed content-related features alongside language accuracy, including grammar and spelling (e.g., Bae and Bachman 2010), with some studies also incorporating additional dimensions such as text length and coherence (e.g., Bae and Lee 2012, which included content, grammar, spelling, text length, and coherence), and organization (e.g., Peterson et al. 2004, which assessed ideas [content], language, organization, and conventions). A review of standardized assessments designed for young or lower-proficiency learners revealed a similar pattern. For example, the Test of English as a Foreign Language (TOEFL) Primary Test emphasizes content, language, and organization, while WIDA writing performance descriptors highlight discourse-level features, grammatical control, and lexical precision across grades. Likewise, the Cambridge English Writing Assessment Scale (B1) includes content, communicative achievement, organization, and language. Across these sources, content-related meaning expression and language use emerged as the most consistently referenced dimensions, while surface-level accuracy, particularly spelling, was also frequently mandated in standards for children's language development (Bae and Bachman 2010).

Second, an analysis of writing assessment criteria presented in 13 middle school English textbooks (Grades 1–3) showed variation in how writing was evaluated across textbooks and grade levels. Despite this variability, several common tendencies were observed. Most textbooks relied on a similar core set of evaluation criteria, with both content and language use included in 12 out of 13 textbooks, whereas spelling or mechanical accuracy appeared less consistently and its inclusion declined across grade levels. Specifically, among Grade 1 textbooks, five included spelling (or spelling and punctuation) as an explicit evaluation criterion; this number declined to three textbooks in Grade 2 and two textbooks in Grade 3. Another observation was the limited attention to organizational features. Across all three grade levels, only one textbook explicitly included organization as an assessment criterion. Overall, the textbook analysis suggests that content and language-related expression form the primary evaluative focus in middle school writing tasks, while spelling and organizational features are treated as secondary components.

Finally, writing-related achievement standards in the expression domain of the 2022 Revised National English Curriculum (Grades 1–3) were examined to ensure alignment between the rating scale and curricular expectations. These standards repeatedly emphasize learners' ability to express experiences, feelings, and ideas clearly on familiar topics and to complete writing tasks in ways that align with communicative purposes. These curricular emphases reinforced the importance of content relevance and task fulfillment, alongside appropriate language use, as central components.

Overall, the document analysis indicates that content and language use (including grammar and vocabulary) constitute the core dimensions of middle school English writing across research, instructional materials, and curriculum standards. Although spelling was not uniformly emphasized across textbooks, its inclusion in prior research and child language standards supported its initial incorporation as a separate criterion in the draft rating scale. Accordingly, three criteria—Content, Language, and Spelling—were selected for the preliminary version of the rating scale.

4.1.2 Evidence from empirical writing samples

To further refine the rating scale, student writing samples were examined to identify recurring performance

features that distinguished adjacent proficiency levels. This stage focused on how the preliminary criteria manifested in actual learner texts.

For the Content criterion, the empirical review revealed clear qualitative differences in how students fulfilled task requirements. Higher-level responses (Levels A–B) not only included all required elements of the prompt but also provided elaborated details and concrete explanations that supported ideas or described experiences vividly. In contrast, mid-level responses (Level C) typically included the required content but did so in a fragmentary manner, often listing ideas without sufficient explanation (e.g., stating what happened without clarifying *why* it was meaningful). Lower-level responses (Levels D–E) frequently omitted key task conditions, included minimal development, or introduced partially off-topic information. These observations led to refinements of the descriptors to more explicitly reference task fulfillment, completeness of required elements, and the degree of elaboration and support.

For the Language Use criterion, distinctions emerged in terms of sentence complexity, grammatical control, and lexical range. At higher levels, students demonstrated attempts at more complex syntactic structures (e.g., subordinate clauses, infinitival constructions) and used varied vocabulary to enrich descriptions, while minor errors were present. Mid-level performances were characterized by mostly complete but shorter and structurally simpler sentences, with noticeable tense inconsistencies or repetitive lexical choices. At lower levels, writing consisted predominantly of short, formulaic, or prompt-recycled sentences, with frequent morphosyntactic errors that reduced clarity. Based on these patterns, the descriptors were refined to emphasize not only accuracy but also the range of language forms, degree of syntactic complexity, and the extent to which errors interfered with meaning.

Importantly, this refinement process resulted in descriptors that reflected observable differences in actual student performance rather than abstract proficiency labels. The final descriptors were thus anchored in empirically identified features that raters could attend to during scoring. A detailed summary of the features identified and their corresponding descriptors is provided in Appendix A.

4.1.3 Evidence from in-service teacher focus group

The draft rating scale was reviewed by three in-service middle school English teachers and one researcher specializing in diagnostic feedback tool development. The review focused on descriptor clarity, level distinctiveness, interpretability of key terminology, and overall usability in classroom-based assessment.

Overall, reviewers agreed that the scale appropriately reflected key dimensions of middle school EFL writing and that the benchmark samples were consistent with the corresponding level descriptors. They noted that the descriptors, which were grounded in observable performance features, facilitated practical differentiation between adjacent levels and were feasible for classroom application.

Feedback primarily concerned clarifying certain terms. For example, in the Language Use criterion, the phrase “uses rich and vivid vocabulary” was considered somewhat abstract, and reviewers suggested incorporating illustrative examples to clarify expectations at higher levels. Similarly, at Level D, the term “basic vocabulary” was viewed as potentially ambiguous and in need of clearer operational guidance. Additional suggestions addressed the consistency of phrasing across levels and the transparency of distinctions between adjacent categories. These comments resulted in minor revisions to descriptor wording and the inclusion of illustrative examples in the rating scale, with more detailed clarification incorporated into the training materials. Overall, the convergence between expert judgments and the proposed descriptors provides additional support for the domain definition inference, indicating that the rating criteria were aligned with key assessment dimensions of middle school EFL writing and interpretable by experienced practitioners.

4.2 Rater Consistency and Severity in Applying the Rating Scale (RQ2)

To address Research Question 2 on the extent to which raters applied the rating scale consistently and without substantial severity differences, inter-rater reliability was examined using K-alpha, followed by an MFRM analysis of rater performance. Because the rating scale assigns categorical levels (A–F), all ratings were converted to numerical values (A = 5 to F = 0) for statistical analysis.

4.2.1 Inter-rater reliability (Krippendorff's alpha)

Inter-rater reliability, as indexed by K-alpha, was examined to obtain an initial estimate of rating consistency. Following Hayes and Krippendorff (2007), coefficients above .80 indicate high reliability, values between .67 and .80 moderate reliability, and values below .67 low reliability. As shown in Table 5, Task 1 yielded moderate agreement for the Content (α = .79) and Language (α = .77) criteria, whereas the Spelling criterion showed low agreement (α = .64). A similar pattern emerged for Task 2: Content demonstrated moderate agreement (α = .68), Language showed relatively high agreement (α = .79), and Spelling again exhibited low agreement (α = .61). Overall, these results indicate that raters achieved generally acceptable agreement for the Content and Language criteria across both tasks, although agreement for Content in Task 2 was relatively modest. Agreement for the Spelling criterion remained consistently low.

Table 5. Krippendorff's Alpha Values

Task	Criterion	K-alpha	Reliability
1	Content	0.79	Moderate
	Language	0.77	Moderate
	Spelling	0.64	Low
2	Content	0.68	Moderate
	Language	0.79	Moderate
	Spelling	0.61	Low

4.2.2 Rater fit and severity differences (MFRM)

Before conducting the MFRM analysis, descriptive statistics of the rating data were examined to understand overall score patterns. As shown in Table 6, the Content and Language criteria displayed similar mean scores (M = 2.89 and 2.71, respectively), whereas the Spelling criterion showed a considerably higher mean score (M = 3.96). The percentile values showed that more than half of the students received the maximum score (5) for Spelling, suggesting a pronounced ceiling effect and limited discrimination for this criterion. By comparison, the distributions for Content and Language were more balanced and showed greater score variability.

Table 6. Descriptive Statistics by Criterion

	Number of ratings	Mean	*SD*	Min	25%	50%	75%	Max
Content	1189	2.89	1.35	1	2	3	4	5
Language	1189	2.71	1.30	1	2	3	4	5
Spelling	1189	3.96	1.38	1	3	5	5	5

Note. Content and Language ratings were assigned on a five-level scale (A = 5, B = 4, C = 3, D = 2, E = 1). Spelling was scored on a three-level scale (AB = 5, CD = 3, E = 1).

Because the K-alpha indicated consistently low agreement on the Spelling criterion, the MFRM analysis was conducted in two stages. Model 1 included all three criteria (Content, Language, Spelling), and Model 2 excluded the Spelling criterion to examine rater performance based on the two criteria that showed acceptable reliability (Content and Language). The specifications of the two models, including their purposes and facet structures, are summarized in Table 7.

Table 7. Specifications of the MFRM Models

Model	Purpose	Facets
Model 1	Analysis conducted to determine whether the Spelling criterion could be retained in the scoring system	examinee (n = 315), rater (n = 5), task (n = 2), criterion (n = 3)
Model 2	Analysis based on the Content and Language criteria after excluding the Spelling criterion	examinee (n = 315), rater (n = 5), task (n = 2), criterion (n = 2)

4.2.2.1 Preliminary analysis including all three criteria (Model 1)

The results of the preliminary MFRM analysis using all three criteria (Model 1) are illustrated in the Wright Map (Figure 3). A Wright map displays the locations of examinees, raters, and other facets along the same scale called a *logit*, enabling a visual comparison of their relative difficulty or severity. In this part of the results, we focus specifically on the fifth column representing the criterion facet, which indicates the relative difficulty of each criterion. The results showed that raters applied the scale most severely for Language, followed by Content, whereas Spelling was scored far more leniently. This indicates that the difficulty of the Spelling criterion was substantially lower than that of the other two criteria.

A closer examination of the criterion facet supported this interpretation (Table 8). Infit and outfit indices summarize how well the ratings for each criterion follow a consistent pattern and whether the scale operates as intended across performances (Eckes 2015). All three criteria showed acceptable infit and outfit values within the recommended range of 0.5–1.5 (Wright and Linacre 1994), indicating that the criteria functioned coherently and did not produce erratic scoring patterns. However, the severity estimates revealed pronounced differences across criteria. As shown in Table 8, the Spelling criterion was located more than two logits below the other two criteria (a difference of 2.98 and 2.54 logits from Language and Content, respectively). A separation of this magnitude corresponds to more than one full step on the rating scale, indicating that Spelling was scored far more leniently than the other criteria and that most performances received high Spelling scores regardless of overall writing quality. In addition, the criterion facet produced an extremely high separation index (24.34), strata value (32.79), and reliability of 1.00, indicating that the three criteria were highly distinguishable from one another and did not operate at a comparable level. Together with the low inter-rater reliability observed for the Spelling criterion in the K-alpha analysis, these results suggest that Spelling contributed little to distinguishing performance levels and was not functioning effectively as part of the rating system. Accordingly, the Spelling criterion was removed from further analysis, and all subsequent MFRM analyses (Model 2) were conducted on the two remaining criteria. The full version of the final rating scale is provided in Appendix B.

Table 8. Criterion Measurement Report from Model 1

Criterion	Measure	*SE*	Infit MnSq	Outfit MnSq
Language	1.14	.04	0.88	0.89
Content	0.70	.04	1.06	1.03
Spelling	-1.84	.07	1.08	1.07
Separation: 24.34, Strata: 32.79, Reliability: 1.00				

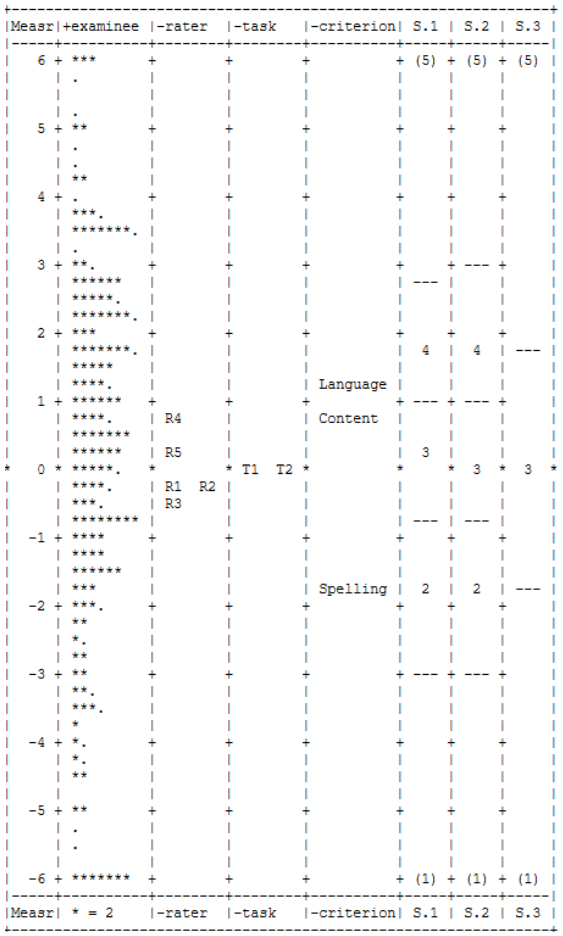

Figure 3. Wright Map from Model 1

4.2.2.2 Main analysis with Content and Language criteria (Model 2)

Rater performance was examined in greater depth using Model 2, which incorporated the Content and Language criteria. The analysis focused on two aspects of rater behavior: (a) whether any rater scored notably more harshly or leniently than the others, and (b) whether the raters applied the scale in a stable and predictable manner.

Figure 4 presents the Wright map for Model 2, illustrating the distribution of examinees, raters, and scoring criteria on a common logit scale. As shown in the second column for the examinee facet, student performances were relatively evenly distributed around the mean, indicating that the writing tasks elicited a suitable range of responses. Although task difficulty and criterion difficulty were not the primary focus of this research question, they are briefly noted for context. The two tasks (T1 and T2) appeared comparable in difficulty, and the Language criterion was slightly more demanding than Content, although the difference was small.

As displayed in the third column for rater facet (Figure 4), rater severity patterns were generally stable. Rater 4 (R4) appeared the most severe, while Rater 3 (R3) was the most lenient. However, the distance between the most severe and most lenient rater was modest, indicating that the raters' scores were broadly aligned and that no rater deviated substantially from the overall severity pattern.

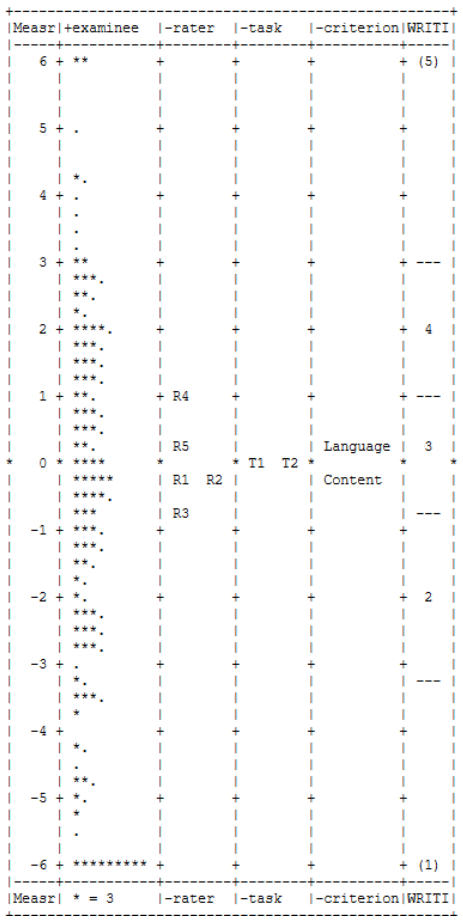

Figure 4. Wright Map from Model 2

Table 9 reports the results for the rater facet, providing more detailed evidence of severity differences and internal rating consistency. Rater severity estimates (i.e., measure) ranged from −0.68 logits (R3; most lenient) to 0.92 logits (R4; most severe), yielding a total span of 1.60 logits. The spread of examinee ability estimates was 13.20 logits. Based on commonly accepted guidelines, rater severity differences are considered acceptable when they fall below one-fourth of this range, which in this case is approximately 3.30 logits (Eckes 2012). The observed rater severity range of 1.60 logits is well below this threshold, indicating that variability in rater severity was modest relative to differences in student performance and within an acceptable range.

Table 9. Rater Measurement Report from Model 2

Raters	Measure	*SE*	Infit MnSq	Outfit MnSq
R4	0.92	.07	1.02	1.03
R5	0.33	.07	1.03	1.04
R1	-0.20	.07	1.24	1.17
R2	-0.37	.07	0.96	0.99
R3	-0.68	.08	0.70	0.71

In terms of internal rating consistency, all raters demonstrated acceptable infit and outfit values within the recommended range of 0.5–1.5 (Wright and Linacre 1994) (Table 9), indicating that raters applied the scale in a

predictable and stable manner. Because none of the raters demonstrated misfit, overall rater reliability was considered satisfactory.

4.3 Functioning of the Rating Scale Categories (RQ3)

Research Question 3 was investigated from two perspectives using the MFRM results from Model 2 to examine whether the rating scale functioned as intended in distinguishing different levels of student performance. Two aspects were examined: (a) how appropriately the two criteria operated within the measurement framework and (b) how the scale categories (i.e., Levels A to E) functioned.

4.3.1 Criterion facet difficulty and fit

The MFRM results of the criterion facet showed that the two criteria behaved in a generally similar and stable manner. First, the measurement report showed only minor differences in criterion difficulty (Table 10). As illustrated in the Wright map (Figure 4), Language was slightly more difficult than Content. The difficulty estimates (as shown in 'measure' in Table 10) ranged from −0.22 logits for Content (less difficult) to 0.22 logits for Language (more difficult), indicating that examinees needed marginally stronger performance to receive higher ratings on the Language criterion. However, the difference was only 0.44 logits. Given the much larger spread of examinee ability (13.20 logits), this difference is minimal and does not suggest any meaningful imbalance between the two criteria (Eckes 2012).

Table 10. Criterion Measurement Report from Model 2

Criterion	Measure	*SE*	Infit MnSq	Outfit MnSq
Language	0.22	.05	0.91	0.94
Content	-0.22	.05	1.08	1.03

Next, the fit statistics indicated that neither criterion showed any signs of misfit. As shown in Table 10, the infit and outfit values for both Language and Content fell within the commonly accepted range of 0.5–1.5, suggesting that the scoring associated with each criterion was stable and coherent. In practical terms, this means that ratings assigned on each criterion followed expected patterns and did not display unpredictable scoring behaviors.

Overall, the criterion-level results indicate that the Content and Language criteria operated comparably and consistently, providing a stable basis for evaluating the functioning of the scale categories in the next section.

4.3.2 Rating scale category functioning

This section examines how well the score categories (Levels A-E) functioned, focusing on whether the distinctions between the adjacent categories were meaningful and whether the overall category structure met commonly accepted criteria for a well-operating rating scale (Eckes 2015). Rating scale statistics from the MFRM analysis (Model 2) were used to evaluate the functioning of the categories.

Previous research has emphasized that a rating scale should meaningfully separate examinees into distinct proficiency levels (e.g., Knoch 2009, McNamara 1996). In line with these guidelines, the examinee facet was first inspected to determine how many performance levels were empirically distinguished by the scale. The examinee strata value was 5.34 (excluding extremes), indicating that examinees were separated into approximately five

distinct levels. This closely aligns with the five-level structure of the rating scale and provides initial evidence that the rating scale differentiated student performance effectively.

Table 11 presents the category statistics from the MFRM analysis (Model 2), including category frequencies, average measures, fit values, and category threshold estimates. These statistics were interpreted based on widely used guidelines for evaluating rating scale quality (Eckes 2015, p. 117).

Table 11. Rating Scale Category Statistics from Model 2

Score	Counts used	%	Average measure	Outfit MnSq	Rasch-Andrich thresholds measure	*SE*
1 (Level E)	357	16%	-3.38	1.1		
2 (Level D)	558	25%	-1.66	1.0	-2.98	.08
3 (Level C)	574	26%	0.07	0.9	-0.77	.07
4 (Level B)	418	19%	1.62	0.9	1.15	.07
5 (Level A)	283	13%	2.79	1.2	2.59	.08

Note. Each threshold marks the boundary between two adjacent categories.

According to this framework, a well-functioning set of score categories should meet six expectations. Table 12 summarizes these indicators along with an evaluation of whether each expectation was met in the present data. Drawing on this evidence, the five-category scale met all of Eckes' (2015) recommended criteria for a well-functioning analytic rating scale.

Table 12. Rating Scale Quality Indicators

Indicator	Higher scale quality	Lower scale quality	Current data
Number of responses per category	$N \geq 10$	$N \leq 10$	$N \geq 10$
Response frequency across categories	Regular	Irregular (skewed, unobserved category)	Regular
Average measures by category	Monotonic increase	Reversals	Monotonic increase
Model fit of rating scale	$MS_U < 2.0$	$MS_U \geq 2.0$	Good fit
Threshold order	Monotonic increase	Disordered	Monotonic increase
Size of threshold increase (logits)	≥ 1.4 and < 5.0	<1.4; ≥ 5.0	≥ 1.4 and < 5.0

Note. MS_U = mean-square outfit statistic.

- All categories were used in meaningful proportions, showing no signs of irregular use, such as highly skewed distributions or patterns suggesting that certain categories were avoided by raters.
- The average measures increased in a clear monotonic pattern across categories. Higher categories consistently corresponded to higher levels of observed writing performance, demonstrating that the ordering of score levels aligned with actual differences in student proficiency.
- Fit indices for the categories fell within acceptable limits, indicating that category use was predictable and coherent.
- The thresholds between adjacent categories were ordered correctly, showing that raters could reliably distinguish neighboring score levels and that each step on the scale represented a meaningful increase in writing quality.
- The distances between adjacent thresholds fell within an appropriate range, suggesting that each transition between categories represented a meaningful step on the scale. Thresholds were neither too close (which would indicate categories that are indistinguishable) nor excessively far apart (which would suggest overly large jumps between score levels).

Figure 5 presents the probability curves for the five score categories. A probability curve illustrates the likelihood that an examinee with a given ability level will receive a particular category. The x-axis represents the underlying writing ability measured in logits, and the y-axis represents the probability of a category being assigned at each ability level. In a well-functioning category, the curves typically show distinct peaks and a logical progression across the ability continuum.

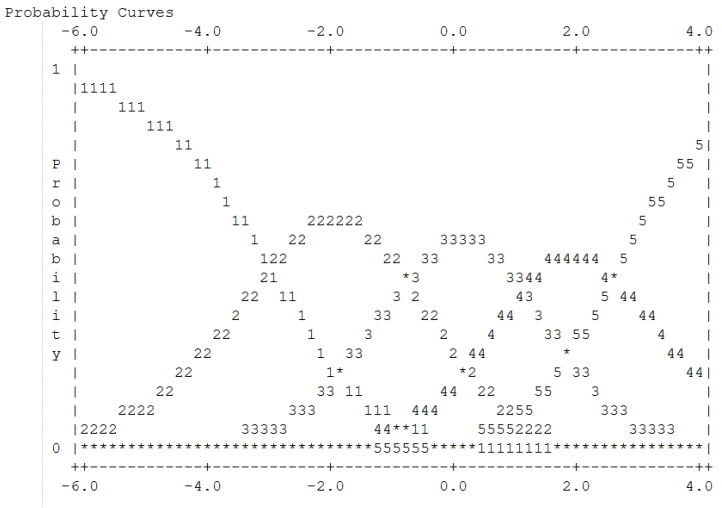

Figure 5. Probability Curve

As shown in Figure 5, the curves for the five categories (1–5 indicating Levels E-A) increase in a generally smooth and orderly manner as examinee ability increases, indicating that the scale operates coherently across the ability continuum. Categories 1 (Level E) and 2 (Level D) cover relatively wide portions of the ability continuum, suggesting that a larger proportion of examinees fall within these lower performance levels and that the scale differentiates performance meaningfully in this range. In contrast, the curves for Categories 4 (Level B) and 5 (Level A) show some overlap in certain regions, indicating that the distinction between these higher performance levels may be somewhat less pronounced. However, the categories maintain identifiable peaks and remain largely separated, suggesting that distinctions at the upper levels are still functioning adequately. Overall, the probability curves demonstrate that the five score categories are ordered appropriately and differentiate examinee performance effectively. No category collapsing or revision appears necessary, providing further support for the validity of the proposed rating scale.

5. Discussion and Conclusion

The present study examined the development and validation of a curriculum-based rating scale for assessing middle school EFL learners' writing within an argument-based validation framework (Knoch and Chapelle 2018). The study focused on two key inferences relevant to rating scale use in rater-mediated assessment: the domain definition and the evaluation inferences. The findings shed light on both the representativeness of the rating scale and the functioning of the scale in actual rating contexts. Taken together, the results support a final rating scale

consisting of two criteria (Content and Language Use), each described across five performance levels. The full version of the scale is provided in Appendix B.

The results related to Research Question 1 provide evidence supporting the domain definition inference, indicating that the proposed rating scale reflects key dimensions of middle school EFL writing. An important aspect of the scale development process is that the construct representation is informed by multiple complementary sources of evidence. Previous work in language testing has emphasized that rating scale development should draw on both test-external sources, such as curriculum documents and instructional materials, and test-internal sources, such as analyses of learner performances (Fulcher 1987, 2003, 2012). The present study incorporated both types of evidence during the development process. First, analyses of the national curriculum and widely used middle school textbooks identified recurring expectations related to the expression of meaningful content and the use of appropriate language forms, ensuring that the rating scale reflected the curricular context in which middle school writing is taught and assessed. Second, analyses of learner writing collected during the pilot phase were used to refine the descriptors. Student texts were examined to identify features that distinguished adjacent proficiency levels, including differences in task fulfillment, elaboration, syntactic complexity, and the extent to which language errors affected clarity. These features informed revisions to the descriptor wording so that the final descriptors captured observable characteristics of learner writing. Incorporating systematic analyses of authentic learner writing therefore helped ensure that the descriptors correspond to performance features that raters are likely to encounter when evaluating middle school learners' texts.

Grounding the descriptors in actual learner production also has implications for the later use of the scale in technology-mediated assessment contexts. The broader project within which the present study is situated aims to develop an AI-assisted feedback tool for middle school English writing. When rating scale descriptors are formulated based on empirically observed features of learner texts, the resulting scale provides clearer reference points for interpreting automated scores and linking them to specific aspects of writing performance. This, in turn, allows feedback to be more closely aligned with learners' actual developmental levels, making it more accessible and meaningful for students, particularly those with low proficiency, as feedback becomes more closely tailored to their developmental level. Previous work on AWE has similarly emphasized that score interpretations become more meaningful when scoring models can be connected to identifiable textual features that correspond to human rating criteria (Attali and Burstein 2006, Wilson and Roscoe 2020). In this sense, the use of authentic learner data during scale development not only strengthens the construct representation of the rating scale but also facilitates the interpretation of scores generated by AI-based systems and supports the generation of actionable feedback for learners and teachers in future applications.

Feedback from the teacher focus group provided additional support for the interpretability of the rating scale. The reviewers generally agreed that the criteria reflected key aspects of classroom writing and that the level distinctions were understandable for practical use. Their suggestions mainly involved minor clarifications to descriptor wording to improve transparency between adjacent levels. Overall, these findings suggest that the rating scale represents middle school EFL writing ability in a way that aligns with curricular expectations, instructional practices, and observed learner performance, supporting the warrant underlying the domain definition inference.

The findings related to Research Questions 2 and 3 address the evaluation inference, which concerns whether the rating procedures operated as intended when evaluating student writing. This inference was examined by analyzing two aspects of the scoring process: rater performance and the functioning of the rating scale categories in distinguishing levels of writing performance. The combined results from the inter-rater reliability analysis and the MFRM analysis provide evidence supporting the evaluation inference, indicating that raters generally applied the rating scale in a reasonably consistent manner. The inter-rater reliability analysis showed moderate agreement

for the Content and Language criteria, whereas the Spelling criterion produced consistently low agreement. The MFRM analysis provided further insight into rater behavior. Overall, raters demonstrated acceptable fit, suggesting that the scale was applied in a stable and predictable manner. In addition, the range between the most severe and most lenient rater remained within commonly accepted limits for rater variability (Eckes 2012), indicating that these differences were unlikely to meaningfully affect the interpretation of students' scores.

One notable finding concerns the Spelling criterion, which behaved differently from the other two criteria. Evidence from the reliability analysis and the initial MFRM results indicates that this criterion did not function as intended in the present assessment context. The strong ceiling effect observed in the descriptive statistics and the consistently low inter-rater agreement indicate that spelling accuracy contributed little to differentiating levels of student writing performance. Several factors may explain this pattern. Because the writing tasks were completed on tablets, raters reported difficulty distinguishing genuine spelling proficiency from typing-related errors, reflecting uncertainty in how such errors should be interpreted. In addition, lower-proficiency students often avoided using more complex vocabulary, resulting in relatively few observable spelling errors and consequently inflated Spelling scores. The limited number of scale levels for spelling may have further constrained its ability to capture meaningful variation in performance. These findings indicate that the Spelling criterion may have introduced construct-irrelevant variance rather than capturing meaningful differences in students' writing ability. The results therefore raise questions about the usefulness of spelling as a separate scoring criterion in electronically produced writing contexts, where factors such as typing errors may reduce its ability to differentiate student performance. Given that the observed limitations appear to be influenced by a combination of task conditions, learner behavior, and measurement-related factors, spelling may be more appropriately treated as part of broader language use or used diagnostically to provide supplementary information, rather than as an independent analytic criterion.

The results of Research Question 3 further contribute to the evaluation inference by examining whether the rating scale categories functioned as intended. Overall, the results indicate that the five score levels operated in a coherent and interpretable manner. The category statistics and probability curves showed that the score levels were used in meaningful proportions and displayed a clear monotonic relationship with examinee ability. The ordered category thresholds suggest that raters were able to distinguish neighboring score levels reliably. The examinee distribution across the scale further indicates that the rating scale was able to differentiate multiple levels of student writing ability, suggesting that the five-category structure provided a stable framework for interpreting student performance.

One important contribution of the present study lies in the use of MFRM to investigate both rater behavior and rating scale functioning within a unified measurement framework. Concerns about rater variability have long been highlighted in research on rater-mediated assessment (e.g., Eckes 2015, Knoch 2009). While traditional reliability indices provide an overall estimate of agreement, they offer limited insight into how differences in rater severity, rating consistency, and scale structure influence scoring outcomes. By modeling multiple facets of the assessment process simultaneously, including examinee ability, rater severity, and criterion difficulty, MFRM makes it possible to examine how these components interact within the measurement system (Eckes 2015). In the present study, this approach allowed not only an examination of rater severity and rating consistency, but also a thorough evaluation of whether the rating scale functioned in a coherent and interpretable manner. These findings highlight that careful rating scale design alone does not necessarily ensure appropriate scale functioning. Empirical examination of rater behavior and category structure, particularly through advanced measurement models such as MFRM, is therefore essential for understanding how rating scales operate in practice.

This methodological approach is also important in the broader context of the present project aiming to develop

an AI-assisted feedback tool for middle school English writing. AI-based assessment systems rely on human-scored training data in order to learn meaningful patterns of writing performance (Wilson and Roscoe 2020). By examining rater behavior through MFRM, the present study contributes to the preparation of a methodologically robust training dataset for the development of AI-based feedback tools.

At the same time, the findings of this study should be interpreted in light of several limitations. The rating scale was developed and validated based on data from two public schools in Seoul and two short personal narrative tasks, and thus the construct representation may be limited to this specific context. Future research should examine the applicability of the scale across a wider range of genres, tasks, and learner populations. In addition, the textbook analysis was based solely on evaluation criteria explicitly presented in student textbooks and may not fully capture more detailed or pedagogically elaborated assessment practices reflected in teachers' guides or classroom-based materials. Accordingly, this aspect of the analysis should be interpreted within this scope, and future research could extend this line of inquiry by examining a wider range of assessment sources.

Despite these limitations, this study illustrates how an argument-based validation framework can be used to guide the systematic development and evaluation of a writing rating scale. By combining curriculum-informed scale design with empirical validation based on learner writing data and MFRM analysis, the study examined whether the scale appropriately represented the target construct and functioned as intended when applied by raters. The findings highlight the value of integrating principled scale design with empirical investigation of rater behavior and scale functioning within a coherent validation framework. More broadly, the study provides a practical example of how researchers and practitioners can approach the development and validation of rating scales, particularly when such scales are intended to support emerging applications such as AI-assisted feedback systems.

References

Aluthman, E. S. 2016. The effect of using automated essay evaluation on ESL undergraduate students' writing skill. *International Journal of English Linguistics* 6(5), 54-70.

Attali, Y. and J. Burstein. 2006. Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment* 4(3).

Bae, J. and L. F. Bachman. 2010. An investigation of four writing traits and two tasks across two languages. *Language Testing* 27(2), 213-234.

Bae, J. and Y. S. Lee. 2012. Evaluating the development of children's writing ability in an EFL context. *Language Assessment Quarterly* 9(4), 348-374.

Cotos, E. 2014. *Genre-Based Automated Writing Evaluation for L2 Research Writing: From Design to Evaluation and Enhancement*. London: Palgrave Macmillan.

Eckes, T. 2012. Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly* 9, 270-292.

Eckes, T. 2015. *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments* (2nd rev. and updated ed.). Bern, Switzerland: Peter Lang Verlag.

Fulcher, G. 1987. Tests of oral performance: The need for data-based criteria. *ELT Journal* 41(4), 287-291.

Fulcher, G. 2003. *Testing Second Language Speaking*. London: Routledge.

Fulcher, G. 2012. Scoring performance tests. In G. Fulcher and F. Davidson, eds., *The Routledge Handbook of Language Testing,* 378-392. London: Routledge.

Hayes, A. F. and K. Krippendorff. 2007. Answering the call for a standard reliability measure for coding

data. *Communication Methods and Measures* 1(1), 77-89.
Huang, S. and W. A. Renandya. 2020. Exploring the integration of automated feedback among lower-proficiency EFL learners. *Innovation in Language Learning and Teaching* 14(1), 15-26.
Hudson, J. A. and L. R. Shapiro. 1991. From knowing to telling: The development of children's scripts, stories, and personal narratives. In A. McCabe and C. Peterson, eds., *Developing Narrative Structure,* 89-135. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
Janssen, G., V. Meier and J. Trace. 2015. Building a better rubric: Mixed methods rubric revision. *Assessing Writing* 26, 51-66.
Kane, M. T. 2013. Validating the interpretations and uses of test scores. *Journal of Educational Measurement* 50, 1-73.
Kane, M. T., B. E. Clauser and J. Kane. 2017. A validation framework for credentialing tests. In C. W. Buckendahl and S. Davis-Becker, eds., *Testing in the Professions: Credentialing Policies and Practice*, 20-41. London: Routledge.
Knoch, U. 2009. Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing* 26(2), 275-304.
Knoch, U. and C. A. Chapelle. 2018. Validation of rating processes within an argument-based framework. *Language Testing* 35(4), 477-499.
Knoch, U., B. Deygers and A. Khamboonruang. 2021. Revisiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing* 38(4), 602-626.
Koltovskaia, S. 2020. Student engagement with automated written corrective feedback: A case study of two language learners' experiences with Grammarly. *Computer Assisted Language Learning* 33(5-6), 510-527.
Korean Ministry of Education. 2022. 2022 revised national curriculum: English. Available online at https://ncic.re.kr/
Link, S., M. Mehrzad and M. Rahimi. 2022. Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computers & Education* 181, 104458.
McNamara, T. 1996. *Measuring Second Language Performance*. Oxford, UK: Blackwell.
McNamara, T. 2002. Discourse and assessment. *Annual Review of Applied Linguistics* 22, 221-242.
Mendoza, A. and U. Knoch. 2018. Examining the validity of an analytic rating scale for a Spanish test for academic purposes using the argument-based approach to validation. *Assessing Writing* 35, 41-55.
Messick, S. 1989. Validity. In R. L. Linn, ed., *Educational Measurement,* 3rd ed., 13-103. Washington, DC & New York: American Council on Education & Macmillan.
Montee, M. and M. Malone. 2014. Writing scoring criteria and score reports. In A. Kunnan, ed., *The Companion to Language Assessment,* Vol. 2, 847-859. Oxford, UK: Wiley-Blackwell.
Peterson, S., R. Childs and K. Kennedy. 2004. Written feedback and scoring of sixth-grade girls' and boys' narrative and persuasive writing. *Assessing Writing* 9(2), 160-180.
Shermis, M. D. and J. Burstein. 2003. *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
Stevenson, M. and A. Phakiti. 2019. Automated writing evaluation: Investigating the feedback–revision relationship and learner engagement. *Language Learning & Technology* 23(3), 66-96.
Tian, L. and Y. Zhou. 2020. Learner engagement with automated feedback, peer feedback, and teacher feedback in an online EFL writing context. *System* 91, 102247.
Whang, S. E., Y. Roh and H. Song. 2023. Data collection and quality challenges in deep learning: A data-centric AI perspective. *The VLDB Journal* 32, 791-813.
Wilson, J. and R. D. Roscoe. 2020. Automated writing evaluation and feedback: Multiple metrics of

efficacy. *Journal of Educational Computing Research* 58(1), 87-125.
Woodworth, J. and K. Barkaoui. 2020. Perspectives on using automated writing evaluation systems to provide written corrective feedback in the ESL classroom. *TESL Canada Journal* 37(2), 234-247.
Wright, B. D. and J. M. Linacre. 1994. Reasonable mean-square fit values. *Rasch Measurement Transactions* 8, 370.
Youn, S. Y. 2015. Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing* 32(2), 199-225.

Examples in: English
Applicable Languages: English
Applicable Level: Secondary

Appendix A. Empirically Observed Features Informing Descriptor Refinement (English Translation)

A1. Content Criterion

Levels	Representative features observed in student texts	Reflected in final descriptor as…
A	Fully addressed all required task elements; provided detailed and vivid explanations of experiences or feelings; included concrete support (e.g., explaining why an experience was meaningful rather than simply stating it); no missing task conditions	"richly and fully elaborated description… fully satisfies task conditions… meaning conveyed clearly and precisely"
B	Addressed most required elements; described three preferences or experiences and their reasons in a balanced manner but with less detailed support; occasionally omitted one minor condition	"generally well-developed description… mostly satisfies task conditions… some minor details lacking"
C	Included required elements but development was fragmentary; for example, stated what happened but did not clearly explain why it was enjoyable or meaningful; ideas listed without sufficient elaboration; minor off-topic content occasionally present	"partially elaborated description… generally follows task conditions but with limited development or minor omissions"
D	Omitted key required elements (e.g., missing reasons or incomplete event descriptions); limited development; partially off-topic or loosely connected content; minimal support for ideas	"very limited or partial description… partially satisfies task conditions… important elements missing"
E	Major omissions of required content; minimal or irrelevant information; extremely limited development; largely off-topic or insufficient length	"highly limited description… fails to satisfy most task conditions… severely limited task fulfillment"

A2. Language Use Criterion

Levels	Representative features observed in student texts	Reflected in final descriptor as…
A	Use of more complex syntactic structures with relative accuracy (e.g., *it is very nice to sleep with my dog*); subordinate clauses (e.g., *when I feel sad, my dog helps me*); varied and vivid vocabulary (e.g., *fluffy fur*); tense use generally accurate with only minor errors not affecting meaning	"uses a variety of appropriate vocabulary and language forms accurately… attempts complex structures… minor errors do not impede clarity"
B	Attempts at complex sentences (e.g., *it helps me when I am tired*) with occasional grammatical errors; mostly appropriate tense use; relatively varied vocabulary (e.g., descriptive verbs such as *crunch*), though sometimes slightly awkward	"uses varied vocabulary and language forms with relative accuracy… some errors present but meaning largely clear"
C	Mostly complete but short or structurally simple sentences; frequent tense inconsistencies; repetition of sentence patterns; limited lexical variety; overall meaning generally understandable despite errors	"uses appropriate but often simple language forms… errors may occur (e.g., tense)… overall meaning generally conveyed"
D	Predominantly short and repetitive sentences; frequent morphosyntactic errors (e.g., overuse of *be* verbs, missing sentence elements); limited vocabulary; partial breakdown of clarity	"uses basic vocabulary and simple forms… frequent errors reduce clarity… meaning partially conveyed"
E	Extremely limited vocabulary and structural control; heavy reliance on prompt recycling; pervasive grammatical errors; severely restricted sentence construction; meaning often unclear	"uses very limited vocabulary and language forms… pervasive errors… meaning largely unclear"

Appendix B. Final Rating Scale (English Translation)

Criterion	**Content & Task Fulfillment** **1. Richness of content and specificity of supporting details** **2. Completeness of task fulfillment[a]** [a]*Note*: Rather than mechanically applying the degree to which task conditions are met, greater emphasis should be placed on whether the content is sufficiently and clearly expressed so that the overall meaning is effectively conveyed.					
Levels	**A**	**B**	**C**	**D**	**E**	**F**
Descriptor	**Richly and fully elaborated description** of experiences related to a familiar topic or of the feelings of a person or character, with detailed and well-supported explanations, so that the **meaning is conveyed clearly and effectively**. The response **fully satisfies task conditions** and is completed in a manner appropriate to the communicative situation and purpose.	**Generally well-developed description** of experiences related to a familiar topic or of the feelings of a person or character, with appropriate support, so that the **meaning is conveyed with relative clarity and effectiveness**. The response **mostly satisfies task conditions** and presents the required content in a generally balanced manner, though **some minor details are lacking**, leaving room for improvement in overall completeness.	**Partially elaborated description** of experiences related to a familiar topic or of the feelings of a person or character, with fragmented or limited support, so that the **meaning is conveyed only in a general or approximate way**. The response **generally follows task conditions**, including most required elements, but shows **limited development or minor omissions**, and may contain partially irrelevant content, leaving the overall completeness somewhat limited.	**Very limited or partial description** of experiences related to a familiar topic or of the feelings of a person or character, with very fragmented or minimal support, so that the **meaning is only partially conveyed**. The response **partially satisfies task conditions**, but shows very limited development: **important elements may be omitted**, or a substantial amount of irrelevant content may be included, resulting in limited overall completeness.	**Highly limited description** of experiences related to a familiar topic or of the feelings of a person or character, with minimal or insufficient support, so that the **meaning is only partly conveyed and remains unclear**. The response **fails to satisfy most task conditions**, with many required elements missing or only minimally developed, and may include irrelevant content or insufficient length, resulting in **severely limited task fulfillment**.	**Little or no description** of experiences related to a familiar topic or of the feelings of a person or character, with **no supporting details**, so that **meaning is barely conveyed or not conveyed at all**. The response **does not satisfy task conditions at all**, containing **little or no relevant content**, resulting in **extremely low task fulfillment**.
Note	Providing **richly and fully elaborated content** rather than including only the minimum required task elements.	**Most main content** specified in the instructions and task conditions **included**, with **occasional limited specificity for each element**.	**Omission of specific required element(s)** specified in the instructions or task conditions (e.g., failing to provide a detailed explanation of the reason).	Key elements in the instructions or task conditions **only partially addressed**.	**Very short** response	**Inadequate** response: not scorable

Criterion	**Language Use[a]** **1. Variety and appropriateness of expressions** **2. Variety and accuracy of language forms** [a]*Note*: Pay attention to cases in which the introductory topic sentence appears to reuse a sentence provided in the task question.					
Levels	**A**	**B**	**C**	**D**	**E**	**F**
Descriptor	**Uses a variety of appropriate vocabulary and language forms accurately**, conveying meaning clearly and effectively. The response **attempts complex structures and generally uses them accurately**, with tense forms expressed almost always correctly. **Minor errors may occur but do not impede clarity**, and rich and appropriate vocabulary is used to express ideas vividly.	**Uses varied vocabulary and language forms with relative accuracy**, so that meaning is **largely clear**. The response **attempts complex structures**, though **some errors are present**, and tense forms are generally used appropriately. **Some errors occur occasionally**, but meaning remains largely clear, and the response demonstrates a relatively varied range of vocabulary, though some expressions may be slightly awkward.	**Uses appropriate but often simple language forms**, allowing the **overall meaning to be generally conveyed**. The response contains many short and simple sentences, and **errors may occur** (e.g., tense errors), though the overall meaning remains understandable. Some **sentence patterns or vocabulary may be repeated**.	**Uses basic vocabulary and simple language forms**, so that **meaning is only partially conveyed**. The response consists mostly of short and simple sentences, and **frequent errors (e.g., tense errors or errors in basic language forms) reduce clarity**, making parts of the message unclear. Sentence patterns and vocabulary are generally monotonous and repetitive.	**Uses very limited vocabulary and language forms**, resulting in **very limited communication of meaning**. **Pervasive errors** in the use of basic language forms or very limited ability to construct sentences make the **meaning largely unclear**.	**Uses little or no vocabulary and language forms**, so that **meaning is barely conveyed or not conveyed at all**. The response may consist only of isolated words without forming sentences, or may be written largely in Korean.
Note	**One or two minor errors** in **complex sentence structures (e.g., complex sentences) or tense use** may occur.	**Attempts** complex sentence structures (e.g., complex sentences) and tense use, though **some errors are present**: **uses** varied vocabulary, though **some expressions may be inaccurate or slightly awkward**.	**Some repetition** of sentence patterns or vocabulary: **tense errors present**.	**Errors in basic grammatical forms** (e.g., overuse of *be* verbs, omission of essential sentence elements).	**Pervasive errors in basic grammatical forms** (e.g., overuse of *be* verbs, omission of essential sentence elements): **occasional copying of parts of the task prompt**.	**Inadequate** response: not scorable