PASAA Journal
Volume 65, January – June 2023, 234–262

PASAA

Revision of an Academic English Writing Rubric for a Graduate School Admission Test

Chawin Srisawat, Kornwipa Poonpon*

Khon Kean University, Thailand

**Corresponding author: korpul@kku.ac.th*

Article information

Abstract

This study aimed to revise an academic English writing rubric for a university graduate admission test, following the Triangulation Scale Revision Process (Banerjee et al., 2015) and the Mixed-methods Conceptual Design Model (Janssen et al., 2015). The current scale revision was informed by a corpus analysis of writing responses, multi-facet RASCH model (MFRM) analysis, and a rater discussion. The samples of the study were 134 scored writing responses. Each writing response was rated by two raters. Seven raters were recruited in the rating process, and four of them participated in the subsequent raters' discussion process. Corpus analysis was employed to investigate the lexical profiles—i.e., lexical diversity, lexical density, and response length—and the grammatical profile, using the online Web Vocabulary Profilers Program and Grammarly, respectively. The MFRM analysis was used for identifying the in-depth correlations among three facets in rating: test takers' abilities, raters' severity, and scales. Finally, a three-hour rater discussion was conducted to identify significant features in rating and to examine the extent to which the rubric could be revised. The corpus analyses revealed that the grammatical and lexical profiles were not significantly correlated (*p*<0.05). The language use trait was then separated into grammar and vocabulary traits. The results of the MFRM showed that the original rubric could be revised in terms of score

E-ISSN: 2287-0024

	weighting, use of decimal scores, and raters' severity. Through discussion, consensus was research to weight each trait equally and to only allow integer scores. The rubric was revised following the results of the corpus analyses by separating language use traits into grammar and vocabulary, and the descriptors were revised according to the significant features retrieved from the raters' discussion. The results of the MFRM also suggested the need for rater training.
Keywords	Academic Writing, Writing Test, Writing Rubric, Writing Rubric Revision
APA citation:	Srisawat, C. & Poonpon, K. (2023). Revision of an Academic English Writing Rubric for a Graduate School Admission Test. *PASAA. 65*, 234–262.

1. Introduction

Academic writing is formal writing with educational aims (Mulvaney et al., 2005; Oshima et al., 2007; Weigle, 2002). In academic writing, emphasis is placed on the originality of thinking, the development of ideas, and the logic of the writer (Weigle, 2002). Moreover, formality, organization, the accuracy of grammar, and language use are also emphasized (Oshima et al., 2007). It is considered more formal and well-organized than creative writing and personal writing, which may include slang, abbreviations, and unfinished phrases. There are many types of academic writing such as essays, journals, response papers, stance or position papers, reviews, abstracts and annotations, informative reports, laboratory reports, research reports, observation reports, proposals or prospectuses, action plans, etc. These types of writing are used for completing their specific writing purposes (Mulvaney et al., 2005). Academic writing is often used as an indicator to determine the degree to which students have mastered not just their thinking and reasoning abilities, but also their cognitive abilities (Weigle, 2002).

To assess learners' academic writing ability, writing scales are necessary tools. Writing scales can reflect tasks and functions that test takers can perform, as well as how well they master linguistic qualities such as coherence, cohesion, vocabulary, and grammar (Bachman & Palmer, 1996; Grabe & Kaplan, 1996; Knoch, 2011; Weigle, 2002). Writing scales can represent test takers' performance in one holistic scale or multiple-trait scales, depending on the aims of the scales' users. Holistic scales emphasize the test takers' accomplishment, while analytic scales provide diagnostic information for their writing progress. In terms of scores or score levels included, writing scales can range from 0 to the highest point, which indicates the proficiency level of a native speaker. Typically, there are seven (plus or minus two) bands (Knoch, 2011). The levels or bands are presented in the form of can-do statements, describing what test-takers can accomplish at each level of the writing scale.

To develop writing scales, several development approaches have been suggested. One well-known approach is the measurement-driven approach (Fulcher et al., 2011). Rubric development from the measurement-driven approach relies on specialists, such as theorists, teachers, or raters, to create the descriptors. Although institutions frequently employ this method, its appropriateness has been criticized. The main criticisms are the lack of connection to the actual performances of the test takers and the inappropriateness of the ability to differentiate scores into levels (i.e., scalability) (Fulcher et al., 2011). Another suggested approach is the performance data-driven approach (Fulcher et al., 2011). This scale's construction begins with the collection of performance data, which is then used to determine or characterize the descriptors. This strategy requires the collection of test takers' performance samples for statistical analysis.

The writing rubric in the present study is an academic English analytic writing rubric for an academic English language test (AELT) at a university, to be henceforth called the AELT writing rubric. This rubric has been used to assess graduate students' writing ability in order to determine their suitability for an academic setting where they must be able to publish their work following the university policy. According to the English Language Department (2015), the original AELT writing rubric was adapted from the scales and descriptors of the TOEFL iBT independent writing rubric and was used in an admissions test for the university's Graduate School (GS). Scoring in the AELT writing test is based on topic development, organization, and language use, highlighting the importance of coherent and well-structured essays with appropriate language usage. Analytic scales are provided for raters to use in their scoring as they are more reliable (Hyland, 2014; Knoch, 2011; Weigle, 2002). Each of the three traits —topic development, organization, and language use—is measured by a five-point scale. The scores reported to the test takers, however, are holistic scores where topic development and organization traits are weighted two times higher than the language use trait. This makes the formula as follows: Total score = Topic Development score*8 + Organization score*8 + Language Use score*4. The highest possible converted score is 100 points (i.e., 5*8 + 5*8 + 5*4 = 100). To ensure fair and consistent evaluation of test takers' writing responses, tests are initially marked by two raters, before being evaluated for inter-rater reliability. The acceptable reliability is at least 0.8; thus, if the scores from raters 1 and 2 differ by more than 20 points, a third rater will be added to score discrepant responses. The results are reported in five holistic score bands (i.e., Band 1 for 0-20 points, Band 2 for 21-40 points, Band 3 for 41-60 points, Band 4 for 61-80 points, and Band 5 for 81-100 points). The results of the test are considered as a profile for master's degree and doctoral degree students who are required to get at least Band 3 and Band 4, respectively, to enroll in graduate programs.

The AELT writing rubric was initially developed based on the measurement-driven approach, as no actual performance had been collected at the beginning of the use of the test. This approach runs the risk of misplacing the test takers into inaccurate levels, as criticized in the study of Fulcher et al. (2011). In addition, the TOEFL Independent Writing Rubric that was initially used for creating the AELT rubric was aimed to be holistic, so when using it as an analytic scale, this may not be able to provide accurate and reliable results. Another concern is overlapping or overly broad traits. For example, the descriptor for the topic development trait mentions "well organized and well developed," which overlaps with the organization trait, potentially causing confusion to the raters. The language use trait contains several features including facility in the use of language, syntactic variety, appropriate word choice, idiomaticity, and lexical or grammatical errors, which may also lead to difficulty in rating. Thus, reviewing and modifying the scale may be needed.

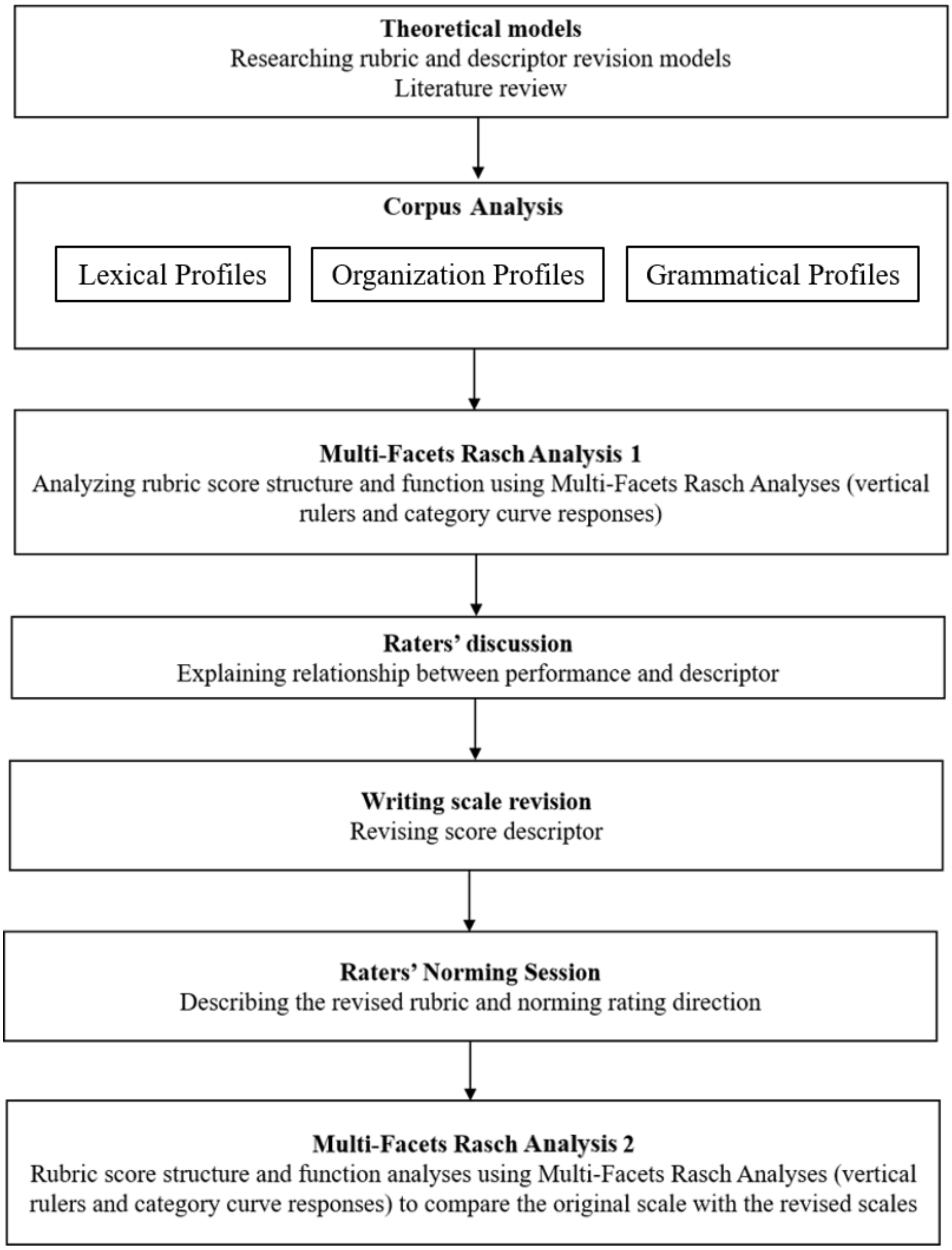
At the time of writing, the AELT has been used as an entrance exam for graduate students for almost ten years, and the results of the test are based on a rubric developed under the measurement-driven approach. Over this time, a significant amount of performance data has been collected, and the need for revision has been observed. Hence, to increase the confidence and reliability in classifying test takers, the performance data-driven approach has been suggested in revising and refining the academic writing rubric. This study aimed to revise the AELT writing rubric by using a revision model adapted from a combination of the triangulation scale revision process (Banerjee et al., 2015) and the mixed-methods conceptual design model (Janssen et al., 2015) in order to evaluate the AELT writing rubric and enhance the reliability and validity of the writing test.

2. The Framework for Writing Rubric Revision

The conceptual framework of this study was shaped by two previous rubric revision models, namely the triangulation scale revision process (Banerjee et al., 2015) and the mixed-methods conceptual design model (Janssen et al., 2015). These two models were adopted in the present study as they are based on the performance data-driven approach in revising writing rubrics in terms of the traits and scales of the rubric and descriptors. The proposed conceptual framework of this study is illustrated in Figure 1. This paper focuses mainly on the first five steps, from Theoretical Models to Writing Scale Revision.

Figure 1

Conceptual Framework of the Current Scale Revision


Theoretical models
Researching rubric and descriptor revision models
Literature review
Corpus Analysis
Lexical Profiles
Organization Profiles
Grammatical Profiles
Multi-Facets Rasch Analysis 1
Analyzing rubric score structure and function using Multi-Facets Rasch Analyses (vertical rulers and category curve responses)
Raters' discussion
Explaining relationship between performance and descriptor
Writing scale revision
Revising score descriptor
Raters' Norming Session
Describing the revised rubric and norming rating direction
Multi-Facets Rasch Analysis 2
Rubric score structure and function analyses using Multi-Facets Rasch Analyses (vertical rulers and category curve responses) to compare the original scale with the revised scales

As shown in Figure 1, the first step of the revision model is Theoretical Models. This step focuses on researching or reviewing aspects concerning writing scale construction or revision, including writing proficiency, writing components, and analysis of linguistic features. In the second step, corpus analysis is used to investigate the characteristics of the writing responses corpus. The focused predictors consist of length, lexical diversity, lexical frequency, cohesion, syntactic complexity, and prompt dependence. The profile analysis is used for explaining the category function by placement level and checking the rubric function across revisions. These steps reveal the significant profiles of each level of the written texts in the writing responses of the target corpus and lead to the decision-making process to revise the rubric. The next step, the multi-facet RASCH Model analysis (MFRM1), is used to analyze the rubric score structure and category function. Based on the mixed-methods conceptual design model (Janssen et al., 2015), the purpose of this step is to investigate the functions of test takers' ability, raters' severity, and scale difficulty. The results from this step can inform the extent to which scales can be revised effectively.

After the components and linguistic features are analyzed, rater discussions can be conducted to assess the usability of the rubric (Banerjee et al., 2015; Janssen et al., 2015). In a rater discussion, the results from the corpus analysis, e.g., lexical diversity and lexical frequency, can be used to inform significant measures in the descriptors for language use (Banerjee et al., 2015). The rater discussion allows raters to explain their scoring process and the correlation between performance and placement and bring up problematic issues and decide on solutions. For example, traits and descriptors can be revised based on the corpus analyses and the raters' discussion. In the triangulated or mixed-method approach, writing scale revision is likely to be carried out and validated effectively (Banerjee et al., 2015).

The main purpose of this paper is to present how the current revised scale was developed based on the corpus analysis, the first MFRM analysis, and the rater discussion. The norming session and the second post-hoc MFMR analysis for scale validation are expected to be done and reported in the future.

3. Methodology

3.1 Writing Responses Used in the Study

The population consists of graduate students who took the writing test. The samples of the study were 134 scored writing responses written by 134 prospective graduate students who took the monthly-offered AELT test. These included responses from each of the five band score levels. The writing test allows test takers one hour for the writing task, after taking the two-hour reading test. They are required to type their responses. The prompt in the writing task is explicit, presenting test takers with a clear question to address in their essay, for example: *"Do you agree or disagree with the following statement? "Soft skills (e.g., communication, interaction, problem-solving) are NOT important at work."* Test takers are required to present and support their arguments on the given topic. The pattern of exposition specified for the test is agreeing or disagreeing, indicating that test takers need to clearly express their stance and provide convincing reasons to support their position. The cognitive demands of the writing test include various skills. Test takers are expected to reproduce facts and ideas accurately, organize and reorganize information effectively, and apply analytical thinking to their arguments. This indicates that the test assesses not only the test takers' linguistic abilities but also how they generate and express their ideas.

For the purpose of score analysis, a linking scoring plan was designed to use seven raters in total to rate these 134 responses. Each of these responses, however, was rated by two trained raters, using the original AELT analytic writing rubric. To ensure interrater reliability of at least 0.80, discrepant responses were rated by a third rater. The scores were divided into five bands, Band 1 as the least proficient and Band 5 as the most proficient. To ensure that the writing topic was the same for all samples, the writing samples were all taken from a single round of the test. The selected round was the round that had the highest number of Band 5 writing responses. Only 14 papers in this round earned Band 5, and all were included in the study. For Bands 1 to 4, 30 samples each were selected by using the random sampling method.

3.2 Raters

Four trained raters who usually rate writing responses for the academic English language test were recruited for this study. These were three female assistant professors and one male lecturer, each of whom held a doctoral degree in TESOL or Applied Linguistics and had more than ten years of experience in English language teaching. They were first contacted via email, in which all the details about their participation were attached. These raters agreed and signed a consent form before participating in a three-hour raters' discussion.

3.3 Procedure

This study used a mixed-methods scale revision approach including quantitative analysis of corpus data and qualitative of the raters' discussion process. The methodology followed the five steps of the model shown in Figure 1: theoretical model, corpus analysis, MFRM analysis, a rater discussion, and scale revision. The explanation for each step is as follows.

Step 1: Theoretical Model

In the first step, the theoretical model used in the present study was identified from the literature review. This includes the appropriate revision model, analysis of lexical and grammatical features, and the rater discussion used for scale revision purposes.

Step 2: Corpus Analysis

The purpose of the corpus analysis was to automatically investigate salient features that appeared in a writing response corpus. The corpus analysis revealed the profiles of the performance at each level according to the original scales. In this study, length, lexical diversity (Type-Token Ratio; TTR), lexical density, Academic Word List (AWL; Coxhead, 2000) families, and errors per sentence unit of the 134 writing responses were analyzed as they were salient features produced by writers with different proficiency levels (Banerjee et al., 2015; Thongyoi & Poonpon, 2020).

Length, lexical diversity, lexical density, and AWL families were analyzed by using the online Web Vocabulary Profiler Program (Cobb, 2002; Heatley et al., 2002). The number of errors per sentence unit was found by using Grammarly to determine the ungrammatical features that appeared in the writing responses, with a sentence boundary being demarcated based on a period. Grammarly was selected for use in this study as it is regarded as the most accurate automated grammar checker that is used as a tool to provide feedback to students' writing (Cavaleri & Dianati, 2016; Ghufron & Rosyida, 2018; O'Neill & Russell, 2019; Qassemzadeh & Soleimani, 2016). Grammarly was able to identify and suggest corrections for grammatical errors ten times more than Word Processing from Microsoft Word (Cavaleri & Dianati, 2016). The length, lexical diversity, lexical density, and the number of errors per sentence unit of each original band were presented in the mean scores and standard deviation. To investigate the actual grammar and lexical abilities of the test takers, the correlations of the length, lexical diversity, lexical density, AWL families, number of errors per sentence unit of each original band, and the score for the language use trait were analyzed by using Pearson's Correlation Coefficient run on IBM SPSS Statistics 19.

Step 3: Multi-Facets RASCH model Analyses

MFRM was employed to analyze scores that were given to the writing responses from the original writing scales. The results from vertical rulers and category curve responses revealed the overlap between levels or bands and further suggested the extent to which the original scales could be revised. In this study, the raw scores of the three traits of the 134 writing responses and the converted scores that were multiplied by the weight of each trait were analyzed. Multi-facet RASCH model analyses were run by using the Facets RASCH model program to create vertical rulers and category response curves to be discussed in the study. Three facets were considered in this study: test takers, raters, and rubric.

Step 4: Raters' Discussion

This step aimed to facilitate an understanding of how raters used the original scales, and elicit their concerns or comments on the original scales. Another objective was to reach a consensus on the extent to which the rubric and descriptors could be revised. The results from the corpus and MFRM analyses were presented in this step, and the raters were given issues to be discussed. In this study, the raters' discussion session was a group discussion of two researchers and four raters. The raters who participated in this study were selected from seven raters who had scored the 134 responses for the MFRM analysis by using a purposive sampling method and a convenience sampling method. Due to time constraints, only four out of seven raters were selected. Since an even number of participants were present in the study, instead of voting and taking the majority's opinions into account, the researchers asked the participants to reach a consensus for any decisions made during the raters' discussion and the revision procedure.

The raters' discussion process consisted of seven steps as follows:

1) The researchers gave instructions.

2) The researchers asked the raters to express their opinion on what features they consider when they rate a written response. Three samples of the writing responses rated by the raters were given to the raters to help them recall their rating process and provide answers to the questions easily.

3) The researchers presented the results of the corpus analyses.

4) The raters discussed and reached a consensus concerning expanding the traits in the rubric and changing the wording in the descriptors based on the significant features they considered.

5) The researchers presented the results of the MFRM.

6) The raters discussed and reached a consensus concerning rating methods.

7) The researchers summarized the participants' ideas for revising the rubric.

Step 5: Scale Revision

This was the subsequent step following the raters' discussion. The rubric and descriptors were revised based on the insights obtained from the discussion forum with the four raters.

4. Results and Discussion

The results of the study were obtained according to the five steps in the developed revision model (Fig. 1). It is noteworthy that the results from each step were used to inform the subsequent steps.

1. The results from the theoretical models

The theoretical model developed for the current study was informed by two theoretical models: the triangulation scale revision process (Banerjee et al., 2015) and the mixed-methods conceptual design model (Janssen et al., 2015). The model used in this study focused on the mixed-methods revision approach, which allows a triangulation scale revision process. This consists of the use of corpus analysis to suggest a profiling methodology that can lead to the expansion of traits, and MFRM analyses to provide a deep understanding of the three facets of test scoring—i.e., test takers' abilities, raters' severity, and scales—and suggest the extent to which the rubric and descriptors can be revised.

2. The results from corpus analysis

The corpus analysis revealed the lexical profiles in terms of length, lexical density, lexical diversity, and grammatical profile (i.e., the number of errors per sentence unit) of the 134 writing responses. Table 1 shows that the longer the writing responses, the higher the bands in which they were placed. Similarly, the more academic vocabulary used in the writing responses, the higher the bands in which they were placed. Higher lexical diversity, in contrast, corresponded with lower band scores. This can be explained by the fact that the greater the total length of the writing responses, the more words were repeated. Lexical density and errors per sentence unit were varied among the five score bands.

Table 1

Lexical profiles and grammatical profiles of the original five bands

Bands	Length		AWL families		Lexical Density		Lexical Diversity		Error per Sentence	
	X^{-}	S.D.	X^{-}	S.D.	X^{-}	S.D.	X^{-}	S.D.	X^{-}	S.D.
1	92.33	79.04	3.70	3.09	0.58	0.11	0.65	0.17	2.15	1.04
2	146.87	54.64	6.90	4.75	0.57	0.07	0.55	0.08	2.97	1.31
3	224.93	41.83	8.10	3.61	0.53	0.06	0.48	0.08	2.83	1.34
4	306.50	53.55	12.97	5.73	0.53	0.04	0.45	0.07	2.76	1.06
5	387.43	105.04	19.50	7.23	0.55	0.03	0.46	0.09	2.26	0.89

In addition, correlation analyses were conducted to examine what features the language use trait was correlated to. The Pearson's correlation coefficient was analyzed (Table 2). The results revealed that the language use score was significantly correlated to length ($p<0.01$), AWL families ($p<0.01$), lexical diversity ($p<0.01$), and lexical density ($p<0.05$). Considering the correlation value, the results of the corpus analyses suggested that test-takers with longer and more complex texts tended to use more sophisticated language, resulting in higher scores in the language use trait. In contrast, the test-takers who used a wider range of vocabulary and more complex grammatical structures may not necessarily have received a higher score in this category. Furthermore, although accuracy must be considered in the language use trait rating, it showed the lowest correlation value with no significant correlation, implying that the number of errors did not significantly affect the overall language use score. According to the results of the correlation analysis, the language use score may reflect the test takers' lexical ability, but it may not reflect their grammatical ability.

Table 2

Pearson's correlation between language use score, lexical profiles, and grammatical profiles

Variables	Length	AWL families	Lexical Diversity	Lexical Density	Errors per Sentence
Language Use Score	.773**	.653**	-.504**	-.204*	-.040

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

According to the results from the corpus analysis, it is possible that the language use trait was rated based on lexical profiles, especially text length, rather than the grammatical profile. This suggests that language use traits should be separated into "grammar" and "vocabulary" and raters should separately rate these two traits individually. This separation of the trait can enhance the reliability of the rubric and descriptors as it allows for more precise analysis (Knoch, 2011; Lee et al. 2008; Weigle, 2002).

Furthermore, both text length and the use of AWL words were found to be crucial factors contributing to higher band scores. Text length played a role in various aspects of the rating process. Firstly, the prompt specified an expected word count of approximately 250 words and test-takers who produced shorter texts than the requirement might have received lower scores. Additionally, the analysis of lexical profiles indicated that longer texts tended to have more word tokens, types, and families, which contributed to a higher score due to the greater vocabulary variety. Moreover, longer texts had the potential to provide more detailed information and elaborate on the test-takers' ideas, showcasing a higher level of language production ability. However, it should be noted that lexical diversity might not effectively reflect the extent of the vocabulary range as mentioned in the rubric and descriptors. As shown in Table 2, longer texts sometimes contained repeated words, resulting in a lower diversity score (TTR).

The previous study of Banerjee et al. (2015), who employed the use of corpus analysis in the rubric revision, also suggested that text length was one variable that can be considered in rating as it was significantly higher in the higher-level writing responses. Other studies have similarly reported that text length was a strong predictor of scores (Weigle, 2002; Wolfe et al., 2016). However, in the present study, the diversity (TTR) decreased slightly as the band level increases, while the density remained relatively constant across all the bands in this study. Thus, the results of the present study contradict the previous studies, in which greater lexical diversity was found to correspond with higher scores. In addition, the present study found that AWL words also play a role in writing ratings as the higher bands showed a higher frequency and higher ratio of AWL words compared to other words.

3. The results of multi-facet RASCH Model analyses

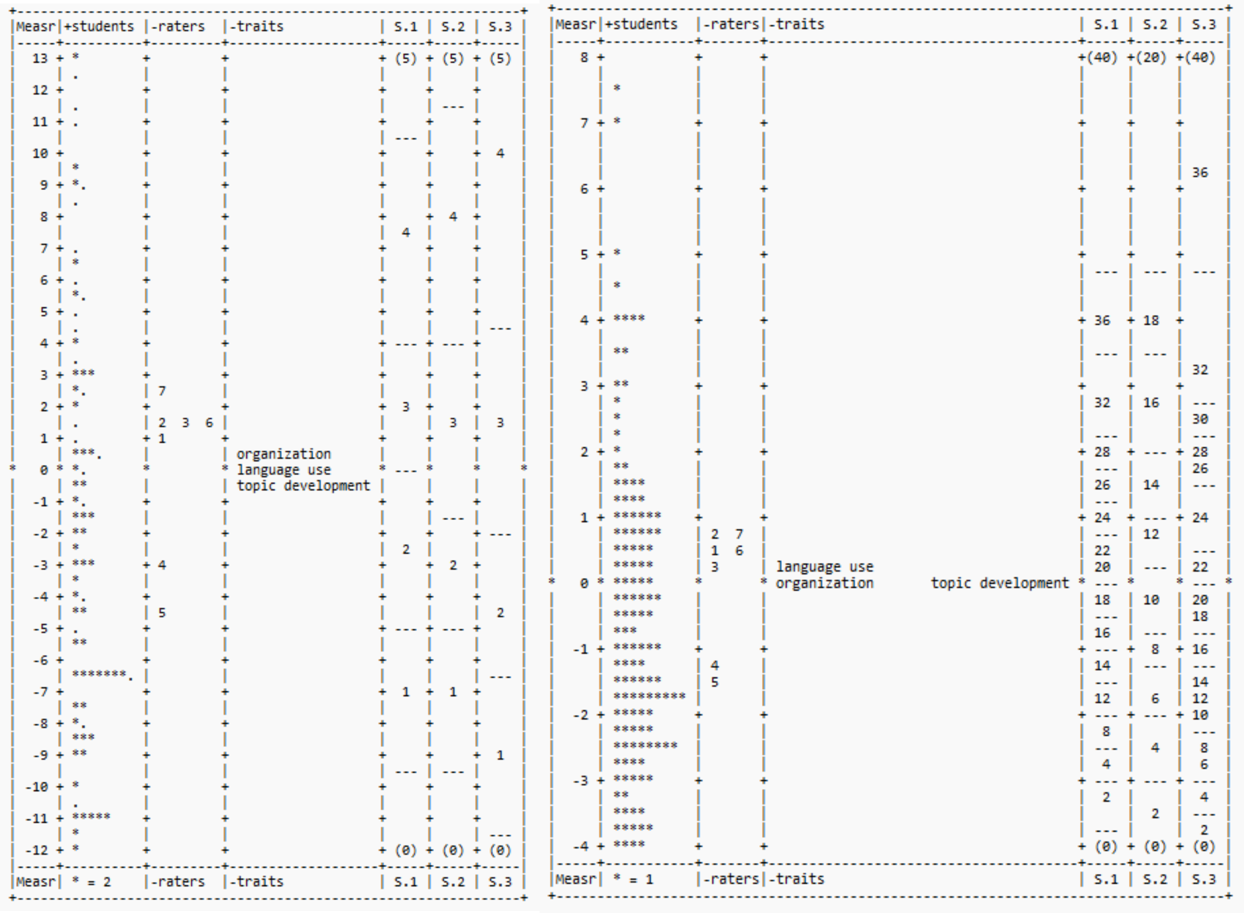
The raw score and the converted score results from the original AELT writing rubric of the 134 rated writing responses were analyzed using MFRM analysis. The common standardized measures were presented as logits in the first column in the vertical ruler (Fig. 2). The results of the raw MFRM score showed that the logit scale centered at 0 and ranged between -12 and 13. The second column shows the test takers' abilities, which varied widely, as the samples were selected from the five different original bands. The reliability of the test takers' model was .96. MFRM is equipped with Infit indices that indicate how well the items fit within the construct, with an appropriate rating being close to 1 or between 0.5 and 1.5 (Linacre, 1994). Out of the 134 responses analyzed, only 58 (43.28%) had Infit indices within the recommended range.

The third column shows the raters' severity. Raters with measures between -1 and 1 severity logits were classified as moderate raters (Linacre, 1994). Raters with measures higher than 1 severity logit were considered severe raters, whereas raters with measures lower than -1 severity logit were considered lenient raters. The reliability of the raters' model was .98. Among the seven raters, only one demonstrated a measure within the range of -1 to 1 (logit = .82) for the raw scores. Meanwhile, four raters exhibited severe scores, with logits of 1.40, 1.55, 1.67, and 2.35. In contrast, two raters demonstrated lenient scores, with logits of -3.07 and -4.74. The results indicated that the seven raters who scored the 134 responses may have had different interpretations of the descriptors when they rated test takers' writing responses.

Interestingly, as shown in the fourth column, the measures for the three traits in the rubric were close to 0. The measures for organization, language use, and topic development traits were .64, -.25, and -.39 logits, respectively. The reliability of the scales' model was .87.

Figure 2

Vertical ruler for the original raw score (left) and vertical ruler for the converted score (right)



```
+------------------------------------------------------------------+
|Measr|+students  |-raters  |-traits             | S.1 | S.2 | S.3 |
|-----+-----------+---------+--------------------+-----+-----+-----|
|  13 + *         +         +                    + (5) + (5) + (5) |
|     | .         |         |                    |     |     |     |
|  12 +           +         +                    +     +     +     |
|     | .         |         |                    |     | --- |     |
|  11 + .         +         +                    +     +     +     |
|     |           |         |                    | --- |     |     |
|  10 +           +         +                    +     +     +  4  |
|     | *         |         |                    |     |     |     |
|   9 + *.        +         +                    +     +     +     |
|     | .         |         |                    |     |     |     |
|   8 +           +         +                    +     +  4  +     |
|     |           |         |                    |  4  |     |     |
|   7 + .         +         +                    +     +     +     |
|     | *         |         |                    |     |     |     |
|   6 + .         +         +                    +     +     +     |
|     | *.        |         |                    |     |     |     |
|   5 + .         +         +                    +     +     +     |
|     | .         |         |                    |     |     | --- |
|   4 + *         +         +                    + --- + --- +     |
|     | .         |         |                    |     |     |     |
|   3 + ***       +         +                    +     +     +     |
|     | *.        | 7       |                    |     |     |     |
|   2 + *         +         +                    +  3  +     +     |
|     | .         | 2  3  6 |                    |     |  3  |  3  |
|   1 + .         + 1       +                    +     +     +     |
|     | ***.      |         | organization       |     |     |     |
*   0 * *.        *         * language use       * --- *     *     *
|     | **        |         | topic development  |     |     |     |
|  -1 + *.        +         +                    +     +     +     |
|     | ***       |         |                    |     | --- |     |
|  -2 + **        +         +                    +     +     + --- |
|     | *         |         |                    |  2  |     |     |
|  -3 + ***       + 4       +                    +     +  2  +     |
|     | *         |         |                    |     |     |     |
|  -4 + *.        +         +                    +     +     +     |
|     | **        | 5       |                    |     |     |  2  |
|  -5 + .         +         +                    + --- + --- +     |
|     | **        |         |                    |     |     |     |
|  -6 +           +         +                    +     +     +     |
|     | *******.  |         |                    |     |     | --- |
|  -7 +           +         +                    +  1  +  1  +     |
|     | **        |         |                    |     |     |     |
|  -8 + *.        +         +                    +     +     +     |
|     | ***       |         |                    |     |     |     |
|  -9 + **        +         +                    +     +     +  1  |
|     |           |         |                    | --- | --- |     |
| -10 + *         +         +                    +     +     +     |
|     | .         |         |                    |     |     |     |
| -11 + *****     +         +                    +     +     +     |
|     | *         |         |                    |     |     | --- |
| -12 + *         +         +                    + (0) + (0) + (0) |
|-----+-----------+---------+--------------------+-----+-----+-----|
|Measr| * = 2     |-raters  |-traits             | S.1 | S.2 | S.3 |
+------------------------------------------------------------------+
```

```
+-----------------------------------------------------------------------------------+
|Measr|+students   |-raters|-traits                               | S.1 | S.2 | S.3 |
|-----+------------+-------+--------------------------------------+-----+-----+-----|
|   8 +            +       +                                      +(40) +(20) +(40) |
|     |            |       |                                      |     |     |     |
|     | *          |       |                                      |     |     |     |
|     |            |       |                                      |     |     |     |
|   7 + *          +       +                                      +     +     +     |
|     |            |       |                                      |     |     |     |
|     |            |       |                                      |     |     |     |
|     |            |       |                                      |     |     | 36  |
|   6 +            +       +                                      +     +     +     |
|     |            |       |                                      |     |     |     |
|     |            |       |                                      |     |     |     |
|     |            |       |                                      |     |     |     |
|   5 + *          +       +                                      +     +     +     |
|     |            |       |                                      | --- | --- | --- |
|     | *          |       |                                      |     |     |     |
|     |            |       |                                      |     |     |     |
|   4 + ****       +       +                                      + 36  + 18  +     |
|     | **         |       |                                      | --- | --- |     |
|     |            |       |                                      |     |     | 32  |
|   3 + **         +       +                                      +     +     +     |
|     | *          |       |                                      | 32  | 16  | --- |
|     | *          |       |                                      |     |     | 30  |
|     | *          |       |                                      | --- |     | --- |
|   2 + *          +       +                                      + 28  + --- + 28  |
|     | **         |       |                                      | --- |     | 26  |
|     | ****       |       |                                      | 26  | 14  | --- |
|     | ****       |       |                                      | --- |     |     |
|   1 + ******     +       +                                      + 24  + --- + 24  |
|     | ******     | 2  7  |                                      | --- | 12  |     |
|     | *****      | 1  6  |                                      | 22  |     | --- |
|     | *****      | 3     | language use                         | 20  | --- | 22  |
*   0 * *****      *       * organization       topic development * --- *     * --- *
|     | ******     |       |                                      | 18  | 10  | 20  |
|     | *****      |       |                                      | --- |     | 18  |
|     | ***        |       |                                      | 16  | --- | --- |
|  -1 + ******     +       +                                      + --- +  8  + 16  |
|     | ****       | 4     |                                      | 14  | --- | --- |
|     | ******     | 5     |                                      | --- |     | 14  |
|     | *********  |       |                                      | 12  |  6  | 12  |
|  -2 + *****      +       +                                      + --- + --- + 10  |
|     | *****      |       |                                      |  8  |     | --- |
|     | ********   |       |                                      | --- |  4  |  8  |
|     | ****       |       |                                      |  4  |     |  6  |
|  -3 + *****      +       +                                      + --- + --- + --- |
|     | **         |       |                                      |  2  |     |  4  |
|     | ****       |       |                                      |     |  2  | --- |
|     | *****      |       |                                      | --- |     |  2  |
|  -4 + ****       +       +                                      + (0) + (0) + (0) |
|-----+------------+-------+--------------------------------------+-----+-----+-----|
|Measr| * = 1      |-raters|-traits                               | S.1 | S.2 | S.3 |
+-----------------------------------------------------------------------------------+
```

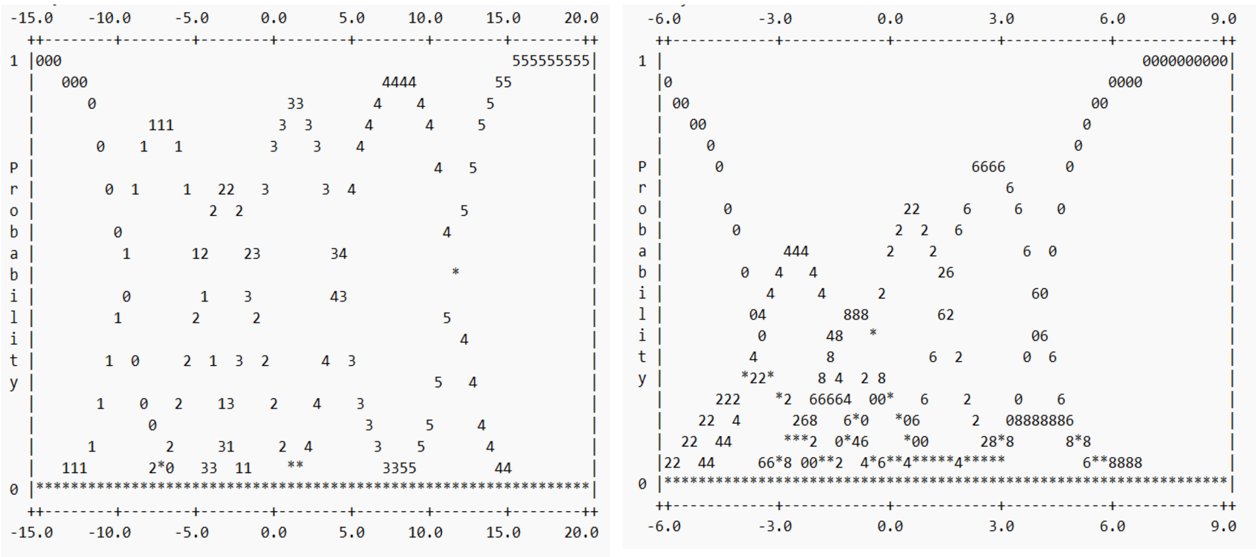
The converted score results of the 134 rated writing responses were also analyzed using MFRM. The results show that the logit scale centered at 0 and ranged between -4 and 8. The second column shows the test takers' abilities, which varied widely and were similar to the results from the raw score. The reliability of the test takers' model was .97. However, the converted score influenced the differences in the Infit indices. There were only 51 test takers (38.06%) whose Infit indices were between 0.5 and 1.5. These findings suggest that the weighting of different traits may have caused misfit scores, resulting in test takers getting scores that did not accurately reflect their abilities.

Regarding raters' severity, the analysis of the converted scores showed that only two out of seven raters had a measure outside the moderate range. The two raters were those whose severity logits were lower than -1 which reflected their leniency in the raw score analyses. The reliability of the raters' model was .99. As the raters rated with raw scores, the raw score MFRM provides a clearer indication of raters' severity compared to the converted score MFRM. The measures of the three traits in the rubric were close to 0. The measures of language use, organization, and topic development traits were .21, -.10, and -.11 logits, respectively. The reliability of the scales' model was .89.

The converted scores may not reflect the actual abilities of the test takers as shown in the category response curves. The comparison of each trait between the raw score MFRM on the left side and the converted score MFRM on the right side shows that the raw score provided clearer thresholds with more clearly defined peaks for each score in all traits. In contrast, the peaks for each scale category in the converted MFRM were overlapping and did not appear in a meaningful order as shown in the example of the category response curves for the language trait (Fig. 3). The category response curves, thus, suggest a problem in converting scores of different traits with different weights.

Figure 3

Category response curves of the raw score MFRM for language use (left) and Category response curves of the converted score MFRM for language use (right)



Another index to be focused on was the Outfit indices of each trait, which reflect how well items fit within the construct. Similar to the Infit indices, the Outfit indices should be close to 1, or between 0.5 and 1.5 (Linacre, 1994). For the raw score MFRM, all traits in all categories showed appropriate Outfit indices except category 5 (5 points of raw score), for which the Outfit indices were .3 for the topic development trait, 2.3 for the language use trait, and .1 in the organization trait. These misfit scores suggest that the actual abilities of the test takers scores may have been misaligned with their scores, which may have been influenced by the severity or leniency of the raters. Outfit indices with misfit logits appeared even more in the converted score MFRM. In the topic development trait, 6 out of 15 score categories exhibited misfit— categories 1, 5, 7, 9, 11, 12, and 14 (i.e., 2, 14, 18, 22, 26, 28, and 36 points converted from .25, 1.75, 2.25, 2.75, 3.25, 3.5, and 4.5 raw scores, respectively). In the language use trait, 2 out of 10 score categories exhibited misfit—categories 1 and 3 (i.e., 2 and 6 points converted from .5 and 1.5 raw scores, respectively). In the organization trait, 3 out of 18 score categories exhibited misfit—categories 2, 5, and 14 (i.e., 4, 10, and 28 points converted from .5, 1.25, and 3.5 raw scores, respectively). The results of the converted score MFRM showed that most of the misfit indices occurred with the score categories that were converted from non-integer scores, suggesting that the use of decimal places may be an inappropriate rating method.

The findings from the raw score MFRM suggest that the seven raters had different severity levels, with most being either too lenient or too severe, resulting in the misalignments between the test takers' scores and their actual abilities. Moreover, the converted score, which was determined by multiplying different weights by the raw scores, resulted in an increased number of misfit measures for the test takers. This suggests three possible aspects of improvement: 1) the need for raters' training sessions to reach a consensus on how they rate writing responses, 2) the reconsideration of how to convert scores by multiplying by different weights, and 3) the change from reporting decimal scores, which contributed to the misfit measures, to giving integer scores.

Another important facet to be considered relates to raters. In this study, raters showed various degrees of severity, from severe to lenient levels, which reflects the authentic situation that occurs in the AELT writing rating. In practice, raters might rate the same writing responses differently; thus, raters' training is further needed to add to this rubric revision model (Bijani, 2011; Khamboonruang, 2023; Schoepp et al., 2018). One reason for this scoring disparity may be that there were only a few rater training sessions provided for the raters beforehand. The MFRM results not only suggest the importance of scale revision, as found by Janssen et al. (2015), but also suggest the importance of raters' training to help all raters enhance their consistency in rating and further enhance the reliability of the rubric and descriptors.

4. The results from the rater discussion

The rater discussion was conducted to present the results from corpus and MFRM analyses to the raters and let them discuss how the rubric and scales could be revised. The raters were asked about the significant features they considered when they rated writing responses. The results are presented in Table 3.

Table 3

Features that were considered in the score rating

Traits	Features to consider
Topic Development	The expected writing responses should - state position - have enough supporting details - express logical ideas - not repeat details across paragraphs - have references - have explanations or exemplifications - contain equal length across all paragraphs
Organization	The expected writing responses should - be in an essay organization: introductory, body, and concluding paragraphs - not be in listing format - be around 250 words in length - use appropriate indentation - use transitional words - show consistency throughout the whole essay
Language Use	The expected writing responses should - be grammatically correct - contain academic or sophisticated vocabulary - contain appropriate mechanics - not have contractions - not be confusing - be well-written

The results from the corpus analysis were presented to suggest the possibility of separating the language use trait into grammar and vocabulary, and the raters agreed with the suggestion. Then, raters were asked the extent to which they believed it was necessary to revise the wording of the rubric. The original

AELT writing rubric and descriptors for all traits were revised following the results of the corpus analyses and raters' discussion. The raters also reached a consensus that zero points would be given to a writing response that merely copies words from the topic, is off-topic, is written in a language other than English (for almost the whole essay), is plagiarized, or is blank.

The raters recommended that the revised descriptors should eliminate any overlapping wording that could cause confusion in rating. For instance, the phrase "*well organized*" in the original descriptors' topic development trait was removed from the rubric since it overlaps with the organization trait. Additionally, the raters incorporated significant features used in rating writing responses into the revised rubric in a clear manner. For instance, the position statement was identified as a critical feature of the topic development trait, and as such, "*explicitly state writer's position*" was determined to be a prerequisite for scores of 3 to 5, while responses that failed to state a position or expressed a questionable would be limited to scores of 1 or 2. Other important features were also included based on the discussions held in the previous step.

The raters also discussed rating methods concerning score weighting and decimal score ratings. The results of the MFRM were presented to identify any concerns regarding these issues. With regard to score weighting, the raters acknowledged that as this test is an English proficiency test, the ability to express ideas through content and produce language should be equally weighted. The original rubric, however, weighted topic development and organization traits twice as heavily as the language use trait. The revised rubric separated the language use trait into grammar and vocabulary traits, resulting in equal weights for all traits. Therefore, there was no need to weight any traits more heavily than others. The raters initially suggested allowing decimal scores to indicate performance between two scale positions, but this idea was rejected since it may not accurately reflect the test taker's ability. For example, the researchers gave the example of a test taker who had been rated 2.5 in all traits and therefore received 50 points in the total converted score; this placed the test taker in Band 3, which the raters agreed that this test taker did not merit. Instead, the raters agreed to weight all revised traits equally and use integer scores, ranging from 1 to 5, without allowing any decimal scores. This would result in a total of 25 points for each trait and a maximum of 100 points for the entire converted score.

5. The results of the writing scale revision

The AELT writing rubric was revised based on the results from the corpus analyses and the raters' discussion (see Appendix). The language use trait was separated into grammar and vocabulary traits. The descriptors were revised following the suggestions in the raters' discussion. The rating methods were revised by disallowing decimal scores and weighting all traits equally. After this step, the raters will be trained to use the revised writing scale and assigned to rate more writing responses. Another MFRM will be needed to validate this revised writing scale. Results from the scale validation will be reported and published in the future.

5. Conclusion

This study proposed a developed scale revision model using corpus analysis together with MFRM analysis. The results from the corpus analysis suggested the possibility of separating traits, alongside other revisions. In this study, the language use trait was suggested to be separated into grammar and vocabulary traits based on the results of the corpus analyses and the raters' discussion. The MFRM yielded in-depth measurements for the test takers' abilities, raters' severity, and scales. The MFRM results revealed three issues to be considered—the different severity or leniency levels of the raters, the score weighting, and the decimal scores. The results of corpus analyses and MFRM were presented to the raters in the raters' discussion process. The decision to separate the language use trait into grammar and vocabulary traits was made and the rubric and descriptors were revised. Finally, the scoring methods concerning score weighting and decimal score allocation were changed. All traits in the revised rubric were weighted equally, and no decimal scores were allowed in the score rating.

This study demonstrates a potential model for revising a writing rubric that may be of benefit to lecturers and test designers. The use of two main analyses—corpus analyses and MFRM—leads to different considerations in scale revision; thus, it can enhance the reliability of the rubric and descriptors from various angles. Separating overly broad traits can enhance the reliability of the rubric and descriptors as it allows for more precise assessment (Knoch, 2011; Lee, 2008; Weigle, 2002). Additionally, this revision process can make the rubric and descriptors more user-friendly as they are revised by the raters themselves. The results of this study showed that, in practice, raters may rate the same writing responses differently; thus, rater training is needed as a further step to this revision model (Bijani, 2011; Khamboonruang, 2023; Schoepp et al., 2018).

After creating or revising a rubric, validation becomes necessary (Janssen et al., 2015; Knoch, 2011; Lee et al., 2008). In addition, Figure 1 proposes two additional steps in the developed revision model: a raters' training session and another MFRM. Therefore, a future study will commence with a raters' training session to ensure consistency in the rating process. Raters will re-evaluate samples of the writing responses, and the rated writing responses will be analyzed using the MFRM to assess the validity of the revised rubric.

The findings of this study offer theoretical, methodological, and practical implications. First, the findings of the study support the use of the revised model developed by combining elements of the triangulation scale revision process (Banerjee et al., 2015) and the mixed-methods conceptual design model (Janssen et al., 2015)—two revision models that suggest revisions from different angles in terms of descriptors and scales. Combining the two models together, the developed model seems to allow for a more comprehensive rubric revision framework than each of the previous models alone. In terms of methodological implications, this study proposed step-by-step procedures, variables, and analysis tools that can be applied in future research. Profiling should be conducted based on the constructs in the rubric; thus, in this study, the language use profiles and the organization profiles were analyzed. This methodology can also be adapted and applied to use with other performance tests, such as speaking tests. Practically, test designers or teachers can use the revised model to develop locally appropriate rubrics from authentic facets. Test designers and lecturers whose test takers have similar characteristics can also use the revised rubric, which was determined to have high reliability and validity, for their testing. For such stakeholders as test takers or score users, this revision model can help to promote confidence that the test scores are likely to precisely place or distinguish test takers' proficiency levels. In the context of the present study, the graduate students who are going to take the AELT can use the descriptors as a guideline to prepare themselves for taking the AELT and meeting the required band as they need.

6. Limitations and Recommendations

The limitations of the study concern the variables used and the number of raters in the raters' discussion. First, two variables in the lexical profiles, namely lexical density and lexical diversity, despite previous findings indicating significant correspondences with writing quality (Banerjee et al., 2015), showed a negative correlation to the language use score in this study. In other words, the lower the lexical density and lexical diversity, the higher the bands in which the test takers were placed. This was confusing when the descriptors expected the higher bands to have a wider range of vocabulary. Therefore, the future study may add more lexical variables into the corpus analysis when profiling the writing responses so they can be used in raters' discussion.

Due to time constraints, only four out of seven raters could participate in the rater discussion. This resulted in using the consensus-building method for decision-making instead of voting. The future study may recruit more raters to gain wider perspectives in the discussion and enhance the reliability of the results of the discussion.

Another recommendation is that the developed rubric revision model should employ the performance data-driven approach (Fulcher et al., 2011); thus, the model may work with speaking test rubric revision as speaking tests are considered performance tests as well. The differences between writing and speaking tests are the task types and the constructs of the rubric, but they are similar in terms of the use of rubrics in score rating. Further studies can use the developed rubric revision to examine the effectiveness of the model in speaking tests.

Further investigations should explore the validity and reliability of the proposed revised writing rubric (Banerjee et al., 2015; Knoch, 2011; Lee, et al., 2008). The same raters or different experienced raters could be used as participants to rate the writing responses. A rater training session should also be given to the raters to ensure their mutual understanding of the revised writing rubric.

7. About the Authors

Chawin Srisawat is a Ph.D. candidate in the Graduate School of Khon Kaen University, Thailand. He obtained his Master’s degree in English from Khon Kaen University, Thailand. Now, he is working as a lecturer in the Faculty of Humanities and Social Sciences, Rajabhat Maha Sarakham University. His current research interests include vocabulary learning and teaching, teaching English writing, writing assessment, the use of technology in language learning, and teachers’ professional development.

Kornwipa Poonpon is an assistant professor in the English Language Department, Faculty of Humanities and Social Sciences, Khon Kaen University, Thailand. She received her Ph.D. in Applied Linguistics from Northern Arizona

University, funded by Fulbright Scholarship. Her research interests include second language assessment, corpus linguistics, EAP and ESP pedagogy, and technology-enhanced language learning.

8. Acknowledgement

Our heartfelt appreciation goes to the AELT raters and individuals who took part in this study, offering their valuable time and personal insights. We extend our thanks to CELEx and their team for their assistance in the preparation and provision of the dataset. Additionally, we are grateful to KKU for generously granting us SPSS license access, exclusively for KKU students. As this paper's first author is in a Ph.D. program, I would like to thank my advisor and all professors for the training and support I have received.

9. References

Bachman, L., & Palmer, A. S. (1996). *Language testing in practice.* Oxford University Press.

Banerjee, J., Yan, X., Chapman, M., & Elliott, H. (2015). Keeping up with the times: Revising and refreshing a rating scale. *Assessing Writing, 26*, 5–19.

Bijani, H. (2011). The effect of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing, 1*(1), 1–16.

Cavaleri, M., & Dianati, S. (2016). You want me to check your grammar again? The usefulness of an online grammar checker as perceived by students. *Journal of Academic Language & Learning, 10*(1), A223–A236.

Cobb, T. (2002).*Web Vocabprofile* [computer program]*.* An adaptation of Heatley, Nation & Coxhead's Range*.* Available at http://www.lextutor.ca/vp/

Coxhead, A. (2000). A new academic word list. *TESOL quarterly, 34*(2) 213–228.

English Language Department. (2015). *Khon Kaen University Academic English Language Test* [Unpublished document]. Faculty of Humanities and Social Sciences, Khon Kaen University.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5–29.

Ghufron, M. A., & Rosyida, F. (2018). The role of Grammarly in assessing English as a foreign language (EFL) Writing. *Lingua Cultura, 12*(4), 395–403.

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing.* Longman.

Heatley, A., Nation, I.S.P. & Coxhead, A. (2002). *RANGE and FREQUENCY programs*. Available at http://www.victoria.ac.nz/lals/staff/paul-nation.aspx.

Hyland, K. (2014). English for Academic Purposes. In Leung, C. & Street, B. (eds.) *The Routledge Companion to English Studies* (pp.392–404). London: Routledge.

Janssen, G., Meier, V., & Trace, J. (2015). Building a better rubric: Mixed methods rubric revision. *Assessing Writing, 26*, 51–66.

Khamboonruang, A. (2023). EFL Thai students' rating performance in self- and peer-assessment of writing: A Many-Facets Rasch analysis. *LEARN Journal, 16*(1), 221–245.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, *16*, 81–96.

Lee, Y., Gentile, C., & Kantor, R. (2008). *Analytic scoring of TOEFL ® CBT essays: Scores from humans and E-rater ®*. (TOEFL Research Report No. RR-81). Educational Testing Service.

Linacre, J. M. (1994). *Many-Facet Rasch Measurement*. MESA Press.

Mulvaney, M. K. & Jolliffe. (2005). *Academic writing: Genres, samples, and resources.* Pearson Education.

O'Neill, R., & Russell, A. M. T. (2019). Grammarly: Help or hindrance? Academic learning advisors' perceptions of an online grammar checker. *Journal of Academic Language & Learning, 13*(1), A88–A107.

Oshima, A., & Hogue, A. (2007). *Introduction to academic writing.* Pearson Education.

Qassemzadeh, A., & Soleimani, H (2016). The impact of feedback provision by Grammarly software and teachers on learning passive structures by Iranian EFL learners. *Theory and Practice in Language Studies, 6*(9), 1884–1894.

Schoepp, K., Danaher, M., & Kranov, A. A. (2018). An effective rubric norming process. *Practical Assessment, Research & Evaluation*, *23*(11), 1–12.

Thongyoi, K., & Poonpon, K. (2020). Phrasal complexity measures as predictors of EFL university students' English academic writing proficiency. *rEFLections*, *27*(1), 44–61.

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

Wolfe, E.W., Song, T., & Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing, 27*, 1–10.

10. Appendix

Revised AELT Rubric and descriptors

Bands	Topic Development	Organization	Grammar	Vocabulary
5	- explicitly states the author's position - effectively addresses the topic and all points fully elaborated - is well developed, using appropriate and sufficient explanations and exemplifications	- displays essay organization (introductory, body, and concluding paragraphs) - displays unity and coherence	- displays a wide range of complex structures -contains almost error-free response -contains no errors that distract the reader	-contains a wide range of academic vocabulary -contains words that are almost never misused and misspelled
4	- explicitly states the author's position - addresses the topic well, though some points may not be fully elaborated - is well-developed, using appropriate explanations and exemplifications	- display essay organization (introductory, body, and concluding paragraphs) - displays unity, and coherence, though it may contain occasional unclear connections	- displays a range of complex structure - contains errors in a few complex constructions - contains errors that may occasionally be distracting	-contains a range of academic vocabulary -contains a few misused and misspelled words
3	- explicitly state the author's position - addresses the topic and task - uses somewhat developed explanations and exemplifications	- display essay organization (introductory, body, and concluding paragraphs) - displays unity, and coherence, though it may contain frequent unclear connections	-displays some complex structure - contains errors in some complex constructions -contains errors that may sometimes be distracting but are not confusing	-contains some academic vocabulary -contains some inappropriate word choices but not causing confusion -contains some misused and misspelled words
2	- questionable author's position - insufficient development in response to the topic and task - insufficient explanations, exemplifications	- incomplete essay organization - inadequate connection of ideas	- displays the majority of simple sentence structure - contains a number of errors in sentence structure -contains errors that are frequently distracting and confusing	-contains a number of basic vocabulary - contains a number of noticeably inappropriate word choices that may cause confusion -contains a number of misused and misspelled words
1	- not state the author's position - limited explanations and exemplifications - serious underdevelopment	- serious disorganization	- serious errors in sentence structure -contains completely distracting and confusing errors	-contains limited vocabulary -contains incorrect word choices that cause confusion -contains a number of misused and misspelled words
0	- merely copies words from the topic, or - is off-topic, or - is written in a foreign language (in almost a whole essay), or - is plagiarizing, or - is blank			