*rEFLections*
Vol 29, No 3, September - December 2022

# Building an Initial Validity Argument for Binary and Analytic Rating Scales for an EFL Classroom Writing Assessment: Evidence from Many-Facets Rasch Measurement

**APICHAT KHAMBOONRUANG**
*Faculty of Humanities and Social Sciences, Mahasarakham University, Thailand*
**Author email: apichat.k@msu.ac.th**

***Article information***

***Article history:***
*Received: 4 Jul 2022*
*Accepted: 6 Dec 2022*
*Available online: 14 Dec 2022*

***Keywords:***
*Validity argument*
*Binary rating scale*
*Analytic rating scale*
*EFL classroom writing assessment*
*Many-facets rasch measurement*

***Abstract***

*Although much research has compared the functioning between analytic and holistic rating scales, little research has compared the functioning of binary rating scales with other types of rating scales. This quantitative study set out to preliminarily and comparatively validate binary and analytic rating scales intended for use in formative assessment and for paragraph writing assessment in a Thai EFL university classroom context. Specifically, this study applied an argument-based validation approach to build an initial validity argument for the rating scales with emphasis on the evaluation, generalization, and explanation inferences, and employed a many-facets Rasch measurement (MFRM) approach to investigate the psychometric functionalities of the rating scales which served as the initial validity evidence for the rating scales. Three trained teacher raters applied the rating scales to rate the same set of 51 opinion paragraphs written by English-major students. The rating scores were analysed following the MFRM psychometrics. Overall, the MFRM results revealed that (1) the rating scales largely generated accurate writing scores, supporting the evaluation inference, (2) the raters were self-consistent in applying the rating scales, contributing to the generalization inference, (3) the rating scales sufficiently captured the defined writing construct, substantiating the explanation inference, and (4) the binary rating scale showed more desirable psychometric properties than the analytic rating scale. The present findings confirm the appropriate functioning and reasonable validity argument of the rating scales and highlight the greater potential of the binary rating scale to mitigate rater inconsistency and cognitive load in a formative classroom assessment.*

## INTRODUCTION

The rating scale plays a central role in the scoring of writing performance since how well raters interpret rating criteria and assign rating scores depends inextricably on the quality of the rating scale (Knoch, Fairbairn & Jin, 2021). A well-developed and validated rating scale not only minimises rating error but also maximises rating quality, hence resulting in meaningful interpretation and use of writing scores as intended in a given context (Knoch & Chapelle, 2018; Knoch, Deygers & Khamboonruang, 2021). Despite the best effort to arrive at a well-

functional rating scale, the characteristics of the rating scale itself (e.g., holistic, analytic, and binary formats) variably influence the rater judgement and score variability (Barkaoui, 2010; Park & Yan, 2019; Wiseman, 2012). In addition, the rater themselves mediate the score variability, exerting various effects (e.g., severity, inconsistency, central tendency, and halo) detrimental to the quality of rating scores (see Myford & Wolfe, 2003). It is therefore of crucial importance to ensure the quality of the rating scale functioning and rater rating performance if the rating score is to be interpreted and used as desired (Knoch & Chapelle, 2018). While no small amount of research has thus far developed and validated rating scales in various writing assessment contexts, much has compared analytic and holistic rating scales (e.g., Barkaoui, 2010, 2011; Ghalib & Al-Hattami, 2015; Harsch & Martin, 2013; Jönsson et al., 2021; Wiseman, 2012), little has developed and validated binary rating scales (Khamboonruang, 2020; Kim, 2010; Lukácsi, 2021; Park & Yan, 2019; Wagner, 2015), and only a paucity has compared the functioning of binary rating scales with rating scales of other types (Jeong, 2019; Park & Yan, 2019). Additionally, although much of the research has applied many-facets Rasch measurement (MFRM) to evaluate the quality of rating scales, very little (e.g., Khamboonruang, 2020; Mendoza & Knoch, 2018) has framed MFRM results within current validation frameworks, in which validity is revisited as to how well evidence supports the proposed interpretations and uses of test scores (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; Kane, 2013, 2021; Messick, 1989, 1995). There is also a call for more research to develop and demystify varying argument-based validation frameworks to be suitable and practical for varying and wider assessment contexts (Kane, 2021; Knoch, 2016). Indeed, MFRM offers a number of rating scale and rater quality indicators that can be used to justify various aspects of current validity which has received very little attention in the Thai EFL language assessment community in particular.

The overall goal of this study is to preliminarily and comparatively validate newly-developed binary and analytic rating scales prior to their operational use in assessing students' paragraph writing ability as part of formative classroom assessment in a Thai EFL university context. In particular, this study applies an argument-based validation approach to construct an initial validity argument for the rating scales with emphasis on the evaluation, generalization, and explanation inferences, and employs a MFRM approach to investigate the psychometric properties of the rating scales which in turn serve as backing for the initial validity argument in question. This study illuminates the functionality and modification of the novel binary and analytic scales and illustrates the alignment of MFRM-based evidence with an argument-based validation framework applied in a local EFL context.

## LITERATURE REVIEW

This section reviews rating scale and rater characteristics in order to highlight their potential effects on language performance assessment and then discusses current validity and validation as well as MFRM which both are adopted as the theoretical frameworks for this research. This section ends with a review of previous research within which the present research is situated.

**Rating scale and rater behaviour**

The rating scale and rater play essential parts in the scoring of writing performance (Knoch, Fairbairn & Jin, 2021; Weigle, 2002). While an effective rating scale can optimise the objectivity of the rater judgement and writing performance scoring, variations in the design features of rating scales can variably affect the rater decision-making behaviour and in turn rating score quality (Knoch & Chapelle, 2018; Knoch, Deygers & Khamboonruang, 2021). To date, holistic, analytic, and binary types of rating scales have typically been used in EFL writing performance assessment. A holistic scale is aimed at evaluating overall performance and requires raters to assign a single score that best represents the overall performance (Weigle, 2002). An analytic scale aims to evaluate different domains of performance and requires raters to assign separate scores to different dimensions of performance and the scores can also be combined to produce a holistic score for overall performance (Weigle, 2002). A binary scale is designed to evaluate very specific attributes or skills on a dichotomous rating category, for instance "yes" or "no", and is typically used for diagnostic purposes (Knoch, Fairbairn & Jin, 2021). In fact, there are variant formats of binary scales observed in the literature. Some look like the Performance Decision Trees scale (see, for example, Fulcher et al., 2011), others are similar to the empirically derived, binary-choice, boundary-definition scale (see, for example, Jeong, 2019; Park & Yan, 2019; Upshur & Turner, 1995 ), and still others take the form of a checklist consisting of multiple yes-no questions or statements representing different language domains (see, for example, Khamboonruang, 2020; Kim, 2010; Lukácsi, 2021; Wagner, 2015). Apart from the rating scale characteristics, differences in rater background and personality mediate performance score variability (Engelhard & Wind, 2018). A well-developed rating scale and well-trained raters notwithstanding, different raters may apply the same rating scale in different manners and exhibit varying effects threatening the validity and fairness of rating scores (Eckes, 2015; Myford & Wolfe, 2003). Amongst the many rater effects, the most problematic effect is severity, in which a rater consistently gives lower or higher scores on average than those given by others (Myford & Wolfe, 2003). Severity is deemed as the most serious and persistent error posing a threat to the scoring validity and is also difficult to minimise (Myford & Wolfe, 2003). All in all, the rating scale functioning and rater rating performance need to be systematically investigated to ensure that the rating scores derived from the rating scale and rater are interpreted and used as intended in a specific context of use.

**Current validity and validation**

To systematically validate a rating scale is to follow current validity and validation concepts. The concepts of validity and validation have been revisited over time. While traditional validity is perceived as the degree to which a test measures what it purports to measure (Chapelle, 2021), current validity is conceptualised as the degree to which test scores are meaningfully interpreted and used as intended by test developers (AERA, APA & NCME, 2014). Of several current validation frameworks proposed by several validity theorists, Kane's (2013) argument-based approach to validation has been widely embraced in language assessment research (Chapelle & Voss, 2021). According to Kane (2013, 2021), validity is a matter of degree, relying on how well the collected evidence complementarily supports the desired interpretations and uses of test scores, and validation is the process of gathering and evaluating evidence to justify

the feasibility of the claims or the intended interpretations and uses of test scores. The argument-based approach involves two sequential and interdependent activities. The first activity is to develop an interpretive/use argument where the intended score interpretations and uses are explicitly articulated through a network of coherent and interconnected inferences, underlying assumptions, and potential backing evidence for the assumptions underlying the inferences. Certain inferences may be more demanding than others and thus require stronger supporting assumptions and hence stronger backing evidence. The inferences typically used in the language assessment discourse include, but not limited to, evaluation, generalization, explanation, extrapolation, decision, and consequence. Readers are highly encouraged to read, for example, Chapelle (2021), Chapelle et al. (2008), Chapelle and Voss (2021), Kane (2013, 2021) and Knoch and Chapelle (2018) for a thorough account of the inferences and the argument-based validation approach. Once an interpretive/use argument is well developed, it then serves as a framework for test development and validity evidence collection. The second activity is to construct a validity argument, in which the collected evidence is evaluated to justify the plausibility of the assumptions underlying the inferences. The inferences that are supported by strong evidence are considered as high validity, whereas those substantiated by weak evidence are regarded as low validity. Researchers may conduct research at a time and formulate research questions aiming to collect certain evidential findings to support certain intended claims (Kane, 2013, 2021).

**Many-facets Rasch measurement**

Many-facets Rasch measurement (MFRM) offers detailed psychometric results that serve as empirical backing for the validity argument in rater-mediated assessments (Eckes, 2015, 2019; Knoch & Chapelle, 2018). Building on Rasch measurement theory (Rasch, 1960), MFRM (Linacre, 1989) is capable of simultaneously examining multiple assessment-specific facets (e.g., rater, examinee, rubric, and task) contributing to measurement variability particularly in a rater-mediated assessment (Linacre, 2022). A MFRM approach simultaneously calibrates raw scores into equally-interval log-odds measurement units (called "*logits*" or "*measures*"), making it possible to compare individual elements within and between facets (Linacre, 2022). The measures represent the estimated parameters of the latent variables associated with the facets under analysis. For example, the latent severity variable is estimated from the rater facet, and the latent ability variable is estimated from the examinee facet. The rater severity measure is adjusted for variations within examinee ability and vice versa. When the observed data fit the expected Rasch model, the measures are exactly on an interval scale and are independent of one another (Eckes, 2015). For instance, the severity measure does not vary according to the varying levels of the examinee ability and vice versa (Linacre, 2022). Accordingly, the measure represents the severity of the rater, the ability of the student, and the difficulty of the criteria more accurately and fairly than raw score-based estimates and is still robust even for missing or incomplete data (Linacre, 2022).

**Previous research**

Previous research has reported different functionalities of different types of rating scales (e.g., holistic, analytic, and binary formats) and their effects on rater behaviours and assessment

outcomes (Barkaoui, 2010, 2011; Harsch & Martin, 2013; Jönsson et al., 2021; Park & Yan, 2019; Wiseman, 2012). For example, raters rated binary descriptors rather consistently (Jeong, 2019; Khamboonruang, 2020; Kim, 2010; Lukácsi, 2021; Wagner, 2015), judged binary descriptors more consistently than analytic scoring criteria (Jeong, 2019; Park & Yan, 2019), and scored easier criteria more congruently than harder criteria (Khamboonruang, 2020). Some raters viewed a binary scale as easy to judge and practical (Jeong, 2019; Khamboonruang, 2020; Park & Yan, 2019) and some perceived that a binary scale was cognitive-loaded and difficult to judge (Kim, 2010; Park & Yan, 2019). In terms of rater behaviours, previous research has discovered that raters' rating variability was influenced by training (Yan & Chuang, 2022), time of rating (Lamprianou et al., 2021), writing genres (Jeong, 2017; Jiuliang, 2014), and rater characteristics, including but not limited to rater experience (Barkaoui, 2010, 2011; Şahan & Razı, 2020), rater fatigue (Mahshanian et al., 2017), rater personality (Zhu et al., 2021), rater age (Isbell, 2017), raters perceptions of criterion importance (Eckes, 2012), and rater styles, strategies and preferences (Han, 2017). Prior studies also found that raters were more consistent in rating higher-quality essays than poorer-quality essays (Han, 2017; Khamboonruang, 2020; Şahan & Razı, 2020), and still significantly differed in their levels of severity even though well-trained and experienced (e.g., Khamboonruang, 2020; Li, 2022; Mendoza & Knoch, 2018; Yan & Chuang, 2022). The review of the previous research indicated that despite much rating scale validation research, very little has compared the functioning between binary and analytic scales and their impacts on assessment outcomes (e.g., Jeong, 2019; Park & Yan, 2019). Only a few studies have been complementarily applied MFRM and validation frameworks to systematically validate rating scales in accordance with current validity and validation concepts (e.g., Khamboonruang, 2020; Mendoza & Knoch, 2018). Only very recently has there been research applying both MFRM and argument-based validation approaches to validate a rating scale in the Thai EFL context (e.g., Khamboonruang, 2020). There is also a call for further elaboration and simplification of an argument-based validation approach to make it more practical to wider assessment contexts (Kane, 2021; Knoch, 2016). Furthermore, the relevant research implicates that despite well-developed rating scales and well-trained raters, rating scale functionalities and rater behaviours always need further investigation to ascertain that rating scores are interpreted and used meaningfully as intended by scale developers. All these provide the rationale for the present research.

## PRESENT RESEARCH

The present research complementarily applies MFRM and argument-based validation approaches to build an initial validity argument for newly-developed binary and analytic scales prior to their operational implementation in a Thai EFL classroom context. The present interpretive/use argument focuses primarily on the evaluation, generalization, and explanation inferences, for which the current research and MFRM analysis could offer evidential backing at this preoperational stage. The evaluation rests on the warrant that the rating scales provide observed scores representative of the student writing performances in the EFL classroom context. This warrant relies on the assumptions that *the rating scales show desirable functioning to ensure accurate rating scores*; and *the raters show desirable performance to ensure accurate rating scores.* The former assumption requires MFRM backing regarding scale functioning accuracy

and the later assumption needs MFRM backing associated with rater performance accuracy. The generalization warrants that the rating scales provides observed scores as estimates of the expected scores across the raters and student writing performances in the EFL classroom context. The assumptions underlying this warrant are that *the rating scales show desirable functioning to ensure consistent rating scores*, requiring MFRM backing related to scale functioning consistency; and *the raters show desirable performance to ensure consistent rating scores*, which needs MFRM backing regarding interrater and intrarater consistency. The explanation inference warrants that the rating scales provides observed scores as estimates of the expected scores attributed to the defined writing construct in the EFL classroom context. This warrant rests on the assumption that *the rating scores are internally consistent with the defined writing construct,* needing MFRM evidence associated with construct coverage. Driven by the interpretive/use argument, this study addresses five research questions aiming to seek MFRM-driven validity evidence to support the assumptions for the proposed inferences:

1) To what extent did the rating scales provide accurate writing scores?
2) To what extent did the raters assign accurate writing scores?
3) To what extent did the rating scales provide consistent writing scores?
4) To what extent did the raters provide consistent writing scores?
5) To what extent did the scale criteria represent the defined writing construct?

## METHODS

### Paragraph writing scripts

In this study, paragraph scripts were written by English-major undergraduates when they took a composition course in 2021 in the Thai EFL classroom context of interest before this study was conducted. A total of 51 opinion paragraphs written by 15 male and 36 female students on the same assignment prompt were used for the current preoperational validation study and were different from those used in the scale development and modification stages.

### Rating scale development

The newly-developed binary and rating scales were designed for assessing English-major undergraduates' paragraph writing ability and the writing scores were intended for making relatively low-stakes formative decisions about teaching and learning improvement in an ongoing Thai EFL writing classroom. The rating scales were designed on the basis of an existing scale, course materials, and teacher intuition which were deemed as relevant and sufficient to inform the new rating scales that would provide meaningful information and use as intended in the context of interest (Knoch, Deygers & Khamboonruang, 2021). The new rating scales were developed and modified over two stages, in which three Thai EFL teachers (including the author) participated in the scale development and modification. All the teacher raters had over three years of EFL writing teaching experience in the context. During the first stage, the author and two female teachers participated in a two-hour session, where we read through an existing scale used in the classroom context and scored the same set of three paragraphs

before we discussed and commented on both the existing scale and paragraphs in a way that would be useful to the development of a new rating scale for classroom assessment purposes. Building on the existing scale, teacher feedback, and course syllabus, the author designed two versions of the draft binary and analytic scales. About two weeks after the first stage, we met again in a two-hour session, in which we independently applied two types of the draft scales to score a new set of three paragraphs. Subsequently, we discussed rating disagreement and further revision of the rating scales. The two scales were again refined before they were used to collect the data in this study. The finalised binary scale (see Appendix A) consisted of 22 rating criteria or descriptors representing seven writing domains similar to those in the analytic scale. Each descriptor was rated on a dichotomous yes-no (1-0) rating category. The finalised analytic scale (see Appendix B) included seven rating criteria which were rated on a three-score (1, 2, 3) rating category.

**Rater rating procedures**

Three Thai EFL teachers, including the author and two female teachers (not the same teachers participating in the scale development and modification stages), participated in the current preoperational stage aiming to preliminarily validate the new rating scales. All the raters were lecturers of English with more than three years of experience in Thai EFL writing teaching and scoring. First of all, we participated in a two-hour rater training session, during which we read through the binary and analytic scales, followed by a further explanation of the criteria by the author. We then scored one paragraph together, followed by a discussion of the rating disagreement. Following this, we independently rated two new paragraphs using both rating scales and then compared the scoring results. Across the two paragraphs and all rating criteria, the average percent interrater agreements were 86% and 81% for the binary and analytic scales, respectively. Finally, we discussed disagreed ratings before independently applying both rating scales at a convenient time to score the same package and ordering of 51 paragraphs on the same opinion prompt. No specific rating guidelines (such as paragraph ordering and scale ordering) were given to the raters. It took about two months for all raters to completely rate all the paragraphs.

**Data analysis**

A three-facet (rater, student, criteria) partial credit many-facets Rasch model was used to analyse the binary and analytic scores separately via the FACETS programme (version 3.84.0; Linacre, 2022). In the binary data analysis, the 22 descriptors were further grouped into seven writing domains. For both datasets, only the student facet was positively oriented and allowed to float along the measure scale. The MFRM results were used to initially examine data-model fit and unidimensionality Rasch requirements to ensure reliable interpretation of the MFRM results, and subsequently investigate the rater performance, student writing ability, and rating scale functioning at the group and individual levels.

## RESULTS

The MFRM results were grouped into three main parts. First of all, the model-data fit and unidimensionality results were presented to ensure meaningful interpretations of the MFRM results. Secondly, a variable map, separation statistics, and fixed chi-squared test were presented to examine overall distributions of the rater severity, student ability, and criterion difficulty for the binary and analytic scales. Finally, individual-level statistics for the binary and analytic scales were presented to demonstrate the rater behaviour, student ability, and rating scale functioning in more detail.

### Data-model fit and unidimensionality

Table 1 presents data-model fit and unidimenationality indicators. Of all the 3,366 binary and 1071 analytic ratings assigned, about 5% and 1% of the unexpected standardized residuals were outside ± 2 and ± 3, respectively, suggesting that the assigned ratings satisfactorily fit the expected ratings generated by the Rasch model (Linacre, 2022). The criterion and rater Infit and Outfit statistics, which can range from 0 to infinity, were generally close to the expected value of 1 and within the acceptable range of 0.50 – 1.50 (Linacre, 2022), implying that the assigned ratings captured the prime dimension of the defined writing construct (Eckes, 2015). The satisfactory data-model fit and unidimensionality thus support the accuracy and independency of the measures computed in this MFRM analysis (Eckes, 2015).

**Table 1**

**Data-model fit and unidimensionality indicators**

| Rasch indicators | Binary scale | Analytic scale |
|---|---|---|
| Total rating assigned by all raters | 3,366 | 1,071 |
| Number of unexpected standardized residuals outside ± 2 | 142 (4.21%) | 53 (4.94%) |
| Number of unexpected standardized residuals outside ± 3 | 40 (1.18%) | 5 (0.46%) |
| Number of criteria showing acceptable fit statistics | 22 (100%) | 7 (100%) |
| Number of raters showing acceptable fit statistics | 3 (100%) | 3 (100%) |

### Overall rater severity, student ability, and criterion difficulty

Figure 1 shows the variable maps of the binary and analytic scales which both display the locations and distributions of the raters' severity, students' ability, and criteria's difficulty measures on the common measure/logit scale in the first column. The higher and lower measures from 0 represent the higher and lower levels of the severity, ability, and difficulty, respectively. Using the binary scale, the three raters (A, B, and C) range in severity measures from -0.52 to 0.41 measures (*M* = 0.35, *SD* = 0.47), with Rater A showing the highest severity. The student ability measures range between -2.03 and 3.85 measures (*M* = 0.37, *SD* = 1.11). The binary descriptors, labelled with their associated writing domains: Main Idea (MI), Supporting Idea (SI), Supporting Details (SD), Concluding Statement (CS), Paragraph Unity (PU), Sentence Use (SU), and Vocabulary Use (VU), range from -2.04 to 3.30 measures (*M* = 0.00, *SD* = 01.26). As for the analytic scale, the map displays a wide range of the rater severity measures (*M* = 0.00, *SD* = 0.61, *Min* = -0.51, *Max* = 0.68), student ability measures (*M* = 0.64, *SD* = 1.00, *Min* = -1.49, *Max* = 3.90), and criterion difficulty measures (*M* = 0.00, *SD* = 1.26, *Min* = -2.04,

*Max* = 3.30). The final column portrays the structure of the analytic score categories (1, 2, 3) assigned across the seven criteria (S.1 – S.7). In general, the three scores were distributed in a desired hierarchical order on the measure scale, where 3, 2, and 1 were placed on the top, in the middle, and at the bottom, respectively. This means that Category 3 was the most difficult to get or represented the highest level of quality and thus requires the highest writing ability (Linacre, 2022). The length of each score represents the proportion of the ratings assigned to each category in each criterion (Linacre, 2022).

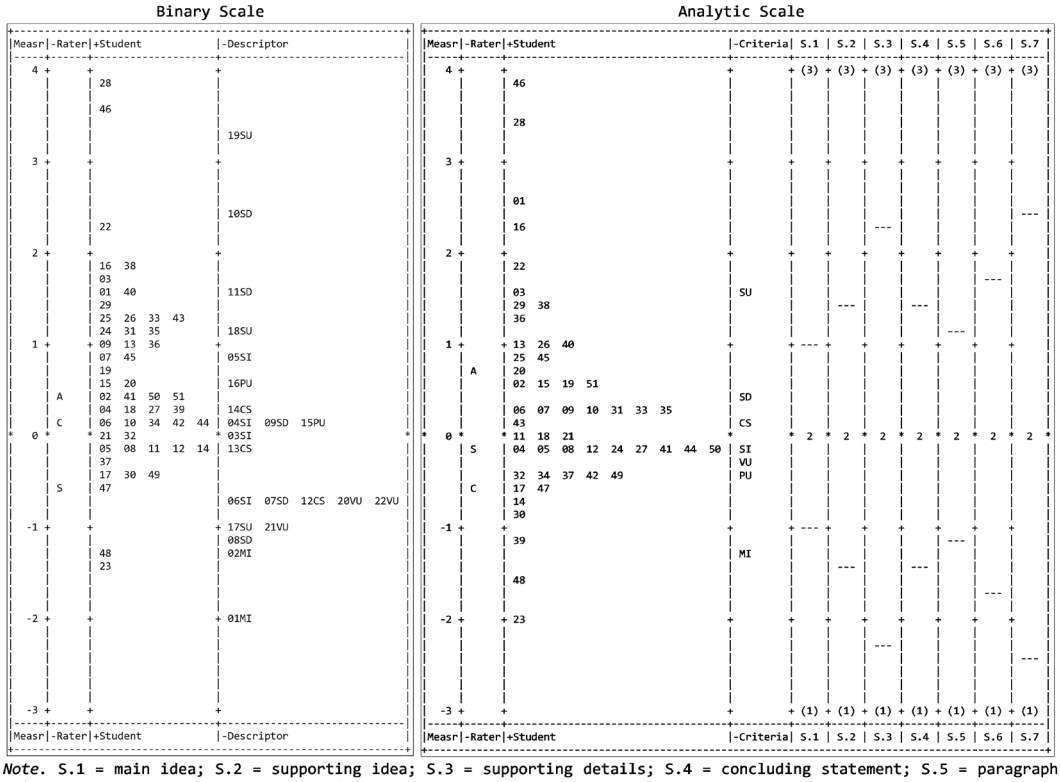


**Figure 1** Variable maps for binary and analytic rating scales

Table 2 shows the fixed chi-squared test and separation statistics of all facets. For both rating scales, the fixed chi-squared tests of all facets were statistically significant ($p < 0.01$) and the separation ratio values, which can range from 0 to infinity, were far greater than the expected value of 1 (Linacre, 2022) for all facets, suggesting that the levels of the rater severity, student ability and criterion difficulty were not homogenous (Linacre, 2022). The rater and criterion separation strata values were around 8, implying that the rater severity and criterion difficulty measures could be stratified into about eight statistically distinguishable levels (Linacre, 2022). The student separation strata index was around 4, which suggests that the student ability measures could be grouped into about four statistically distinct levels. The separation reliability, which can range from 0 to 1, of all facets were about 0.9, indicating that the rater severity, student ability and criterion difficulty measures were highly reliably different (Linacre, 2022). For both rating scales, the rater severity heterogeneity indicates a low degree of interrater reliability between the raters (Eckes, 2019). Nonetheless, none of the raters exhibited a severity

measure outside ±1, suggesting that each rater was not overly severe or lenient (Eckes, 2019). The student ability heterogeneity suggests that the rating scales and raters were sensitive enough to reliably distinguish between high- and low-ability students (Linacre, 2022). The criterion difficulty heterogeneity implies that the student sample was large enough to reliably span the criterion difficulty hierarchy of about eight statistically distinct levels (Linacre, 2022), and that the rating criteria sufficiently represented the defined construct, supporting the construct validity of the rating scales (Linacre, 2022). Comparatively, the binary scale showed relatively wider measure distributions than the analytic scale across all the facets.

**Table 2**

**Separation statistics and fixed chi-square test**

| | Rater severity | | Student ability | | Criteria difficulty | |
|---|---|---|---|---|---|---|
| **Statistics** | **Binary** | **Analytic** | **Binary** | **Analytic** | **Binary** | **Analytic** |
| Fixed chi-squared test | $P < 0.01$ | $p < 0.01$ | $p < 0.01$ | $p < 0.01$ | $p < 0.01$ | $p < 0.01$ |
| Separation ratio | 6.38 | 5.99 | 2.97 | 2.41 | 5.96 | 5.56 |
| Separation strata | 8.84 | 8.32 | 4.30 | 3.55 | 8.28 | 7.74 |
| Separation reliability | 0.98 | 0.97 | 0.90 | 0.85 | 0.97 | 0.97 |

## Rater rating performance

Table 3 presents the rater statistics for the binary and analytic scales. Each rater rated 51 paragraphs on 22 binary descriptors and 7 analytic criteria, thus making up a total number of 1,122 binary and 357 analytic ratings, respectively. Since different raters assigned different scores for each rating, their total rating scores were different. For both scales, the standard error of estimate (SE) values were very close to 0, thus confirming the high precision of the severity measures (Eckes, 2015), with the binary scale showing lower SE values and hence more precise severity measures. Of all the raters, Rater A was the most severe, exercising the highest severity measures of 0.41 and 0.68 for the binary and analytic scales, respectively. The raters Infit and Outfit statistics were all within the acceptable range, implying that each rater accurately and self-consistently assigned ratings across the rating criteria and writing performances (Eckes, 2015). None of the rater Infit and Outfit statistics were also below 0.50 and over 1.50, which indicates that each rater' ratings were not unduly similar and inconsistent, respectively (Eckes, 2015). The acceptable rater fit statistics supports a high degree of intrarater reliability and independent expert rater for each rater (Linacre, 2022). Given the total 3,366 binary ratings, the raters showed the observed interrater agreement of 63.4%, slightly below the expected agreement of 65.9% by 2.5%, indicating a little lower-than-expected agreement between the raters (Linacre, 2022). Of all the 1,071 analytic ratings, the rater showed the 48.6% observed interrater agreement, far below the 53.3% expected agreement by 4.7%, indicating a much lower-than-expected agreement between the raters (Linacre, 2022). Overall, the raters used the binary scale more congruently than the analytic scale.

**Table 3**

**Rater statistics for binary and analytic rating scales**

| Rater statistics | Binary scale | | | Analytic scale | | |
|---|---|---|---|---|---|---|
| | Rater A | Rater C | Rater S | Rater A | Rater S | Rater C |
| Total rating | 1122 | 1122 | 1122 | 357 | 357 | 357 |
| Total score | 613 | 672 | 785 | 690 | 778 | 810 |
| Severity measure | 0.41 | 0.11 | -0.52 | 0.68 | -0.17 | -0.51 |
| Severity SE | 0.07 | 0.07 | 0.08 | 0.10 | 0.10 | 0.10 |
| Infit statistic | 0.88 | 1.13 | 0.96 | 0.89 | 1.23 | 0.92 |
| Outfit statistics | 0.94 | 1.11 | 1.04 | 0.86 | 1.26 | 0.85 |
| Observed interrater agreement | | 63.4% | | | 48.6% | |
| Expected interrater agreement | | 65.9% | | | 53.3% | |

## Student writing ability

Table 4 presents the student statistics for the binary and analytic scales. Due to limited space, only the students showing both Infit and Outfit statistics outside the acceptable bound were presented. Overall, the student ability means were 0.64 and 0.37 for the binary and analytic scales, respectively, which suggests that the binary scale provided comparatively higher ability estimates. The student measures based on the binary and analytic scales showed a strong and positive correlation ($N = 51$, $r = 0.92$, $p < 0.01$), implying that the rating scales and raters were consistent in differentiating the student ability. The students generally showed acceptable fit statistics for the binary scale, which suggests that the binary ratings assigned to the students were in line with the Rasch model (Eckes, 2015). In other words, a fitting student means that the student of a given proficiency had a greater probability of receiving a higher rating from more lenient raters than more severe raters on easier items than on harder items (Eckes, 2015). For the analytic scale however, five students (9.80%) showed both Infit and Outfit statistics outside the acceptable bound and thus were considered as misfitting. This implies that these students were not generally assigned rating scores that were consistent with the levels of the raters' severity and criteria's difficulty (Eckes, 2015). On the whole, the binary scale seemed to provide more consistent ratings than the analytic scale across the raters and paragraphs.

**Table 4**

**Student statistics for binary and analytic rating scales**

| No. | Binary scale | | | | Analytic scale | | | |
|---|---|---|---|---|---|---|---|---|
| | Measure | SE | Infit | Outfit | Measure | SE | Infit | Outfit |
| 04 | 0.32 | 0.29 | 1.14 | 1.17 | -0.14 | 0.39 | 1.74 | 2.20 |
| 05 | -0.16 | 0.28 | 1.19 | 1.29 | -0.14 | 0.39 | 1.64 | 1.96 |
| 19 | 0.65 | 0.30 | 1.05 | 1.00 | 0.50 | 0.41 | 0.41 | 0.37 |
| 43 | 1.32 | 0.33 | 0.97 | 1.28 | 0.18 | 0.40 | 0.40 | 0.37 |
| 51 | 0.48 | 0.29 | 0.83 | 0.73 | 0.50 | 0.41 | 0.40 | 0.35 |

## Rating scale functioning

Table 5 shows the statistics of the binary and analytic criteria associated with Main Idea (MI), Supporting Idea (SI), Supporting Detail (SD), Concluding Statement (CS), Paragraph Unity (PU), Sentence Use (SU), and Vocabulary Use (VU). The criteria were arranged in descending order

of difficulty. For both rating scales, the criterion fit statistics were generally acceptable, suggesting that each criterion was applied consistently across the raters and student paragraphs and the criteria were internally consistent in measuring the defined writing construct (Eckes, 2015). This serves as evidence of the unidimensional measurement and construct validity (Linacre, 2022). The MI-associated criteria were generally harder while the SU-related criteria tended to be easier than the others, implying that this group of Thai EFL students were good at writing the main idea of a paragraph but poor at writing grammatical sentences. Between the two rating scales, the difficulty levels of the criteria associated with MI, SU, SI, and SD were consistent while those related to CS, PU, and VU were slightly different.

**Table 5**

**Binary and analytic criterion statistics**

| Rating scale | Criteria | Measure | SE | Infit | Outfit |
|---|---|---|---|---|---|
| Binary | 19SU | 3.30 | 0.30 | 1.12 | 0.88 |
| | 10SD | 2.42 | 0.23 | 1.05 | 1.04 |
| | 11SD | 1.52 | 0.19 | 1.26 | 1.54 |
| | 18SU | 1.10 | 0.18 | 0.94 | 0.88 |
| | 05SI | 0.84 | 0.18 | 0.93 | 0.89 |
| | 16PU | 0.55 | 0.18 | 0.83 | 0.76 |
| | 14CS | 0.27 | 0.18 | 1.17 | 1.20 |
| | 09SD | 0.17 | 0.18 | 0.88 | 0.82 |
| | 15PU | 0.17 | 0.18 | 0.84 | 0.78 |
| | 04SI | 0.07 | 0.18 | 0.83 | 0.76 |
| | 03SI | -0.03 | 0.18 | 0.97 | 0.92 |
| | 13CS | -0.16 | 0.18 | 0.85 | 0.77 |
| | 20VU | -0.71 | 0.20 | 1.12 | 1.69 |
| | 06SI | -0.75 | 0.20 | 1.13 | 1.18 |
| | 07SD | -0.75 | 0.20 | 0.99 | 1.03 |
| | 12CS | -0.75 | 0.20 | 0.99 | 1.01 |
| | 22VU | -0.75 | 0.20 | 1.21 | 1.28 |
| | 17SU | -0.97 | 0.21 | 1.14 | 1.42 |
| | 21VU | -1.01 | 0.21 | 0.91 | 0.74 |
| | 08SD | -1.21 | 0.22 | 0.85 | 0.95 |
| | 02MI | -1.26 | 0.23 | 1.05 | 1.21 |
| | 01MI | -2.04 | 0.29 | 0.95 | 0.90 |
| | Mean SU | 1.14 | 0.23 | 1.07 | 1.06 |
| | Mean SD | 0.43 | 0.20 | 1.01 | 1.08 |
| | Mean PU | 0.36 | 0.18 | 0.84 | 0.77 |
| | Mean SI | 0.03 | 0.19 | 0.97 | 0.94 |
| | Mean CS | -0.21 | 0.19 | 1.00 | 0.99 |
| | Mean VU | -0.82 | 0.20 | 1.08 | 1.24 |
| | Mean MI | -1.65 | 0.26 | 1.00 | 1.06 |
| Analytic | SU | 1.62 | 0.16 | 1.05 | 1.04 |
| | SD | 0.48 | 0.18 | 0.77 | 0.74 |
| | CS | 0.07 | 0.14 | 0.92 | 0.89 |
| | SI | -0.10 | 0.14 | 0.80 | 0.77 |
| | VU | -0.34 | 0.19 | 1.14 | 1.17 |
| | PU | -0.48 | 0.13 | 0.91 | 0.84 |
| | MI | -1.26 | 0.15 | 1.52 | 1.48 |

Table 6 reports the score category statistics for 22 binary descriptors. Both 0 and 1 scores showed the number of counts (observed ratings) over the required minimum of 10, confirming precise and stable estimates of the score category statistics (Linacre, 2004). Category 1 showed more counts than Category 0, meaning that the raters tended to assign 1 more frequently than 0 across the descriptors (Linacre, 2004). Yet, no descriptors displayed Outfit statistics over the required maximum of 2, suggesting that both 0 and 1 scores were not overly used, nor were they used in an idiosyncratic manner (Linacre, 2004). For all descriptors, 0 and 1 displayed the average (average ability) measures close to the expected measures, with those of the highest score (1) exhibiting higher measures than those of the lowest score (0). This means that the students receiving the highest score on a particular criterion were more proficient than those assigned the lowest score on the same criterion (Linacre, 2004).

**Table 6**

**Binary score category statistics**

| | | | Measure | | |
|---|---|---|---|---|---|
| Descriptor | Score | Count | Average | Expected | Outfit |
| 01MI | 0 | 14 | 1.75 | 1.89 | 0.9 |
| | 1 | 139 | 2.77 | 2.75 | 1.0 |
| 02MI | 0 | 26 | 1.36 | 1.19 | 1.3 |
| | 1 | 127 | 2.00 | 2.04 | 1.0 |
| 03SI | 0 | 57 | 0.09 | 0.13 | 0.9 |
| | 1 | 96 | 1.01 | 0.98 | 1.0 |
| 04SI | 0 | 60 | -0.15 | 0.05 | 0.7 |
| | 1 | 93 | 1.03 | 0.90 | 0.9 |
| 05SI | 0 | 84 | -0.66 | -0.60 | 0.9 |
| | 1 | 69 | 0.36 | 0.29 | 0.9 |
| 06SI | 0 | 37 | 0.95 | 0.75 | 1.2 |
| | 1 | 116 | 1.53 | 1.60 | 1.1 |
| 07SD | 0 | 37 | 0.76 | 0.75 | 1.1 |
| | 1 | 116 | 1.59 | 1.60 | 1.0 |
| 08SD | 0 | 27 | 0.87 | 1.15 | 1.0 |
| | 1 | 126 | 2.05 | 1.99 | 0.9 |
| 09SD | 0 | 63 | -0.17 | -0.04 | 0.8 |
| | 1 | 90 | 0.92 | 0.82 | 0.8 |
| 10SD | 0 | 125 | -1.97 | -1.99 | 1.1 |
| | 1 | 28 | -0.97 | -0.87 | 1.0 |
| 11SD | 0 | 104 | -1.00 | -1.19 | 1.3 |
| | 1 | 49 | -0.61 | -0.22 | 1.6 |
| 12CS | 0 | 37 | 0.76 | 0.75 | 1.0 |
| | 1 | 116 | 1.59 | 1.60 | 1.0 |
| 13CS | 0 | 53 | 0.05 | 0.25 | 0.7 |
| | 1 | 100 | 1.20 | 1.09 | 0.8 |
| 14CS | 0 | 66 | 0.07 | -0.12 | 1.3 |
| | 1 | 87 | 0.60 | 0.74 | 1.1 |
| 15PU | 0 | 63 | -0.22 | -0.04 | 0.8 |
| | 1 | 90 | 0.95 | 0.82 | 0.8 |
| 16PU | 0 | 75 | -0.53 | -0.36 | 0.8 |
| | 1 | 78 | 0.68 | 0.52 | 0.8 |
| 17SU | 0 | 32 | 1.22 | 0.94 | 1.5 |
| | 1 | 121 | 1.71 | 1.78 | 1.1 |

| Descriptor | Score | Count | Measure | | Outfit |
|---|---|---|---|---|---|
| | | | Average | Expected | |
| 18SU | 0 | 92 | -0.88 | -0.83 | 0.9 |
| | 1 | 61 | 0.17 | 0.09 | 0.9 |
| 19SU | 0 | 138 | -2.76 | -2.78 | 1.2 |
| | 1 | 15 | -1.67 | -1.49 | 0.8 |
| 20VU | 0 | 38 | 0.99 | 0.72 | 1.9 |
| | 1 | 115 | 1.47 | 1.56 | 1.1 |
| 21VU | 0 | 31 | 0.78 | 0.98 | 0.7 |
| | 1 | 122 | 1.87 | 1.82 | 0.9 |
| 22VU | 0 | 37 | 1.10 | 0.75 | 1.3 |
| | 1 | 116 | 1.48 | 1.60 | 1.2 |

Table 7 lays out the score category statistics for seven analytic criteria. Overall, 1, 2, and 3 score categories showed acceptable and desirable counts, measures, and Outfit statistics across the criteria. FACETS reported the threshold measure (difficulty measure) only for the analytic rating score categories. As shown in the table, the threshold measures of 1, 2 and 3 increased monotonically from the lowest to the highest score categories, suggesting that the highest score (3) was more difficult to get than the lower (2) and lowest (1) scores (Linacre, 2004). The step advances (the distances between the threshold measures of the adjacent categories) were greater than 1.4 measures, suggesting a sufficient distinction between the score categories (Linacre, 2004). Yet, the step advance (from -0.43 to 0.43 threshold measures) on the MI criterion was below 1.4 measures, signalising that the 2-score category was not sufficiently distinct from other categories or not representative of a distinct level of the MI quality (Linacre, 2004). Yet, all the step advances did not exceed 5 measures, suggesting no significant rating centrality and/or dependency problems for all score categories (Linacre, 2004).

**Table 7**

**Analytic score category statistics**

| Criteria | Score | Count | Measure | | Outfit | Threshold measure |
|---|---|---|---|---|---|---|
| | | | Average | Expected | | |
| MI | 1 | 12 | 0.90 | 0.33 | 1.5 | low |
| | 2 | 37 | 1.43 | 1.05 | 1.4 | -0.43 |
| | 3 | 104 | 1.78 | 1.98 | 1.5 | 0.43 |
| SI | 1 | 26 | -0.73 | -0.49 | 0.8 | low |
| | 2 | 79 | 0.21 | 0.26 | 0.5 | -1.23 |
| | 3 | 48 | 1.53 | 1.53 | 0.8 | 1.23 |
| SD | 1 | 21 | -1.56 | -1.05 | 0.7 | low |
| | 2 | 114 | -0.12 | -0.17 | 0.7 | -2.31 |
| | 3 | 18 | 1.60 | 1.29 | 0.8 | 2.31 |
| CS | 1 | 28 | -0.83 | -0.63 | 0.8 | low |
| | 2 | 82 | 0.19 | 0.13 | 0.7 | -1.29 |
| | 3 | 42 | 1.27 | 1.25 | 1.0 | 1.29 |
| PU | 1 | 23 | -0.49 | -0.20 | 0.8 | low |
| | 2 | 61 | 0.61 | 0.51 | 0.7 | -0.82 |
| | 3 | 69 | 1.49 | 1.49 | 1.0 | 0.82 |
| SU | 1 | 65 | -1.75 | -1.87 | 1.2 | low |
| | 2 | 77 | -1.15 | -1.00 | 1.0 | -1.63 |
| | 3 | 11 | 0.92 | 0.54 | 0.7 | 1.63 |
| VU | 1 | 9 | 0.20 | -0.38 | 1.2 | low |
| | 2 | 115 | 0.52 | 0.51 | 1.3 | -2.48 |
| | 3 | 29 | 1.60 | 1.83 | 1.1 | 2.48 |

Figure 2 portrays the probability curves of the three score categories for seven analytic criteria. The horizontal axis is the student ability measure scale while the vertical axis represents the probability of being rated in each category. Students with higher ability measures had higher probability of being rated in higher categories, while those of lower proficiency had higher probability of being rated in lower categories. Students having ability measures at a threshold measure, where two probability curves of two adjacent categories cross, had a 50/50 probability of being rated in either of the two adjacent categories. As can be seen, the score category probability curves for almost all criteria showed observably separate and distinct peaks, indicating that individual score categories were sufficiently distinct from each other or representative of the distinct levels of the writing ability domains or criteria (Linacre, 2004). Yet, Category 2 of the Main Idea (MI) criterion exhibited a very low and narrowed curve which was largely overlapped by the curves of Categories 1 and 3. This signals that Category 2 was not adequately distinct from other categories or not representative of a unique quality for the Main Idea criterion (Linacre, 2004), which is also consistent with the MI step advance below 1.4 measures.

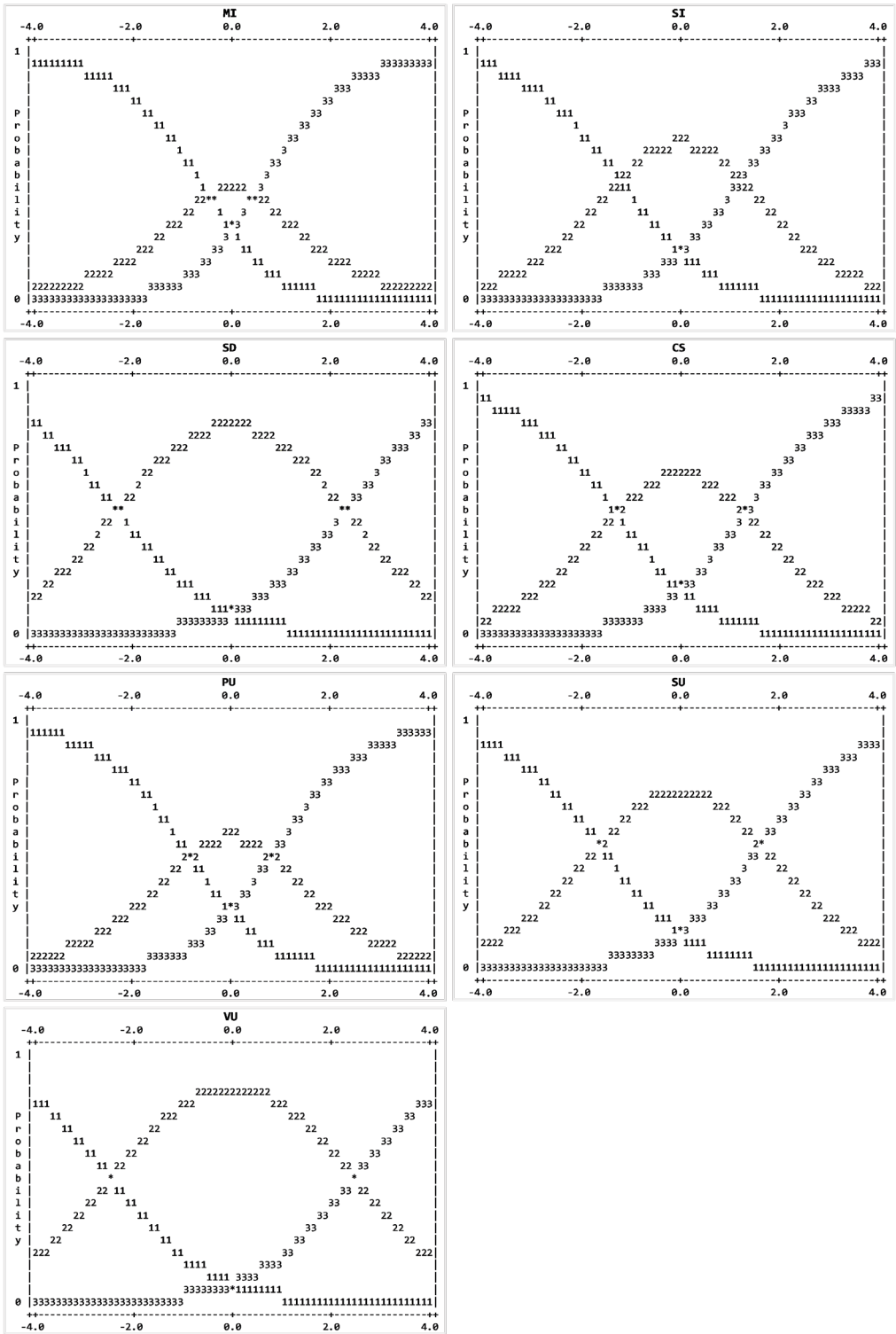


**Figure 2** Rating score probability curves of analytic criteria

## DISCUSSION

This section discusses the key findings in relation to previous research and in alignment with the inferences specified in the current interpretive/use argument. The MFRM evidence for each inference is discussed and evaluated with a view to constructing the initial validity argument for the rating scales. Limitations are also discussed in this section.

Figure 3 displays the validity argument structure portraying the alignment of strong (✓) and weak (✗) MFRM evidence with the assumptions underlying the evaluation, generalization, and explanation inferences. In terms of the evaluation inference, the MFRM results support that the rating scales and raters provided accurate ratings as evidenced by the acceptable data-model fit, unidimensionality, fit statistics, and score category statistics. Yet, the analytic scale showed some unacceptable student fit statistics and an undesirable score category of the Main Idea criterion, partially threatening the analytic rating accuracy. One possible factor underlying the score category problem would be that the descriptions of the score categories might not be sufficiently clear or distinct from each other. Therefore, the descriptions need to be reworded or the score category that was not adequately distinct from others may be deleted or merged with another category (Linacre, 2004). Another underlying cause would be that the raters might have failed to distinguish between the score categories, hence necessitating more substantial rater training (Linacre, 2004). Taken together, it can be argued that there is reasonable MFRM evidence to substantiate the feasibility of the evaluation inference that the rating scales provide observed scores representative of the student writing performances in the EFL classroom context.

With regard to the generalization inference, the MFRM results generally substantiate that the rating scales provided consistent scores across the raters and student performances as evidenced by acceptable criterion and student fit statistics. Despite the severity and agreement variability, each rater was not overly severe or lenient and was self-consistent in applying the rating scales as evidenced by acceptable rater fit statistics. Overall, the raters used the binary scale more consistently than the analytic scale, supporting previous findings (Jeong, 2019; Park & Yan, 2019). Yet, the raters showed inconsistent analytic ratings on some students as suggested by the unacceptable student fit statistics. This thus partially undermines the rating consistency of the analytic scale. The high severity variability and low interrater agreement in this research imply that despite receiving the rater training, the raters still differed in interpreting the criteria, probably resulting from insufficient rater training and/or unclear rating criteria. Other factors that might have influenced the rater variability in this study could be those found in previous research, including rater fatigue (Mahshanian et al., 2017), rater personality (Wiseman, 2012), rater perceptions of criterion importance (Eckes, 2012), rater rating styles, strategies and preferences (Han, 2017), and essay characteristics (Han, 2017; Khamboonruang, 2020; Şahan & Razı, 2020). The high severity variability in this study supports a body of research which revealed that even experienced and well-trained raters still differed in severity (Khamboonruang, 2020; Li, 2022; Mendoza & Knoch, 2018; Yan & Chuang, 2022). Yet, Eckes (2015) pointed out that, in a MFRM analysis, rater severity heterogeneity is not of grave concern since variations in rater severity are adjusted for the estimation of student ability. Instead, it is necessary that each rater maintains rating expertise and is consistently either severe or lenient with respect

to other raters in assigning ratings. In this study, since the rating scales were aimed for use in formative classroom assessment which is rather low-stakes and low-standardised by nature, high interrater agreement or rater severity homogeneity is not necessarily expected in such assessment condition. It can thus be argued that there is reasonable MFRM evidence to support the plausibility of the generalization inference that the rating scales provides observed scores as estimates of the expected scores across the raters and student writing performances in the EFL classroom context.

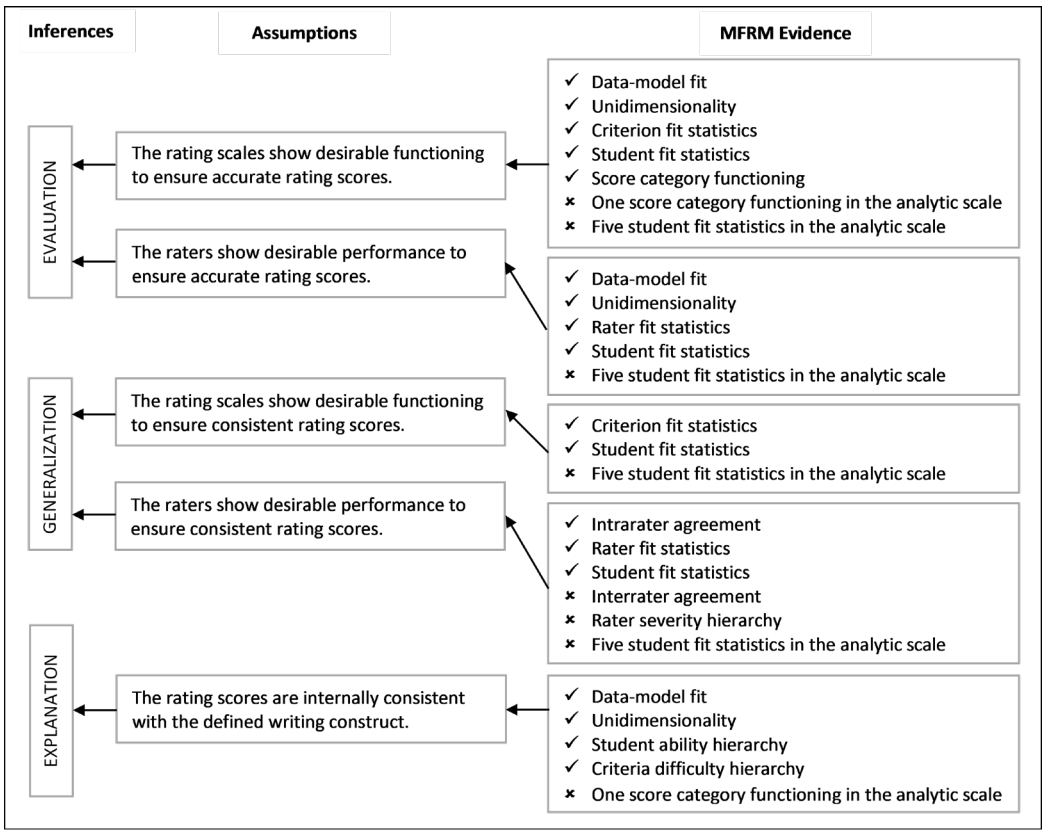


**Figure 3** The validity argument structure

In relation to the explanation inference, the MFRM results overall support that the rating criteria sufficiently captured the defined EFL writing construct as evidenced by the wide spread of the student ability and criteria difficulty measures as well as the data-model fit and unidimensionality evidence. Yet, the analytic scale category on the main idea did not adequately differentiate the skill quality as designed, thereby needing further revision. Based on the measures of the binary and analytic scales, this group of Thai EFL students tended to show persistent problems in the sentence and supporting detail writing skills, but write well on the main idea, probably due to the fact that they were taught clear patterns to structure the main idea which is only one short element of the paragraph, making it possible for the students to receive high ratings on the main idea. Both rating scales showed similar criterion difficulty levels for very difficult criteria (Sentence Use and Supporting Detail) and the easiest criterion (Main Idea) but showed variations in criterion difficulty for not too difficult or easy criteria

(Concluding Statement, Paragraph Unity, and Vocabulary Use). This implies that both rating scales and all the raters were very accurate in assessing very difficult and easy criteria or in other words very weak and strong writing skills. It should be noted that the criteria difficult hierarchy vary according to the characteristics of rating scales, examinees, tasks, and writing genres used in a given assessment context as was revealed in previous research (Jeong, 2017; Jiuliang, 2014). Since this research used written paragraphs on the same writing genre and prompt, it was impossible for the current findings to examine the effect of writing genre and task on the criterion difficulty and student ability estimates. All in all, there is reasonable MFRM evidence to support the feasibility of the explanation inference that the rating scales provides observed scores as estimates of the expected scores attributed to the defined writing construct in the EFL classroom context.

This study is not without limitations which could have influenced the present findings. Due to time constraint, the raters did the rating practice on only two performance samples during the rater training. If the raters had practised scoring more writing performance samples, their rating performances might have been different. Additionally, variations within each rater' rating condition could variedly have influenced the rating scores. However, this is typical in classroom assessment in which teachers normally evaluate students' writing performances under varying conditions. The ordering of paragraph rating and scale use could also have affected the rater performance and rating scores on both rating scales. Yet, it was not known whether the raters followed the same ordering of paragraph rating and scale use or applied the binary scale before the analytic scale and vice versa when scoring each student paragraph. Furthermore, the written paragraphs rated in this study were all in the opinion genre. If the rating scales had been applied to rate other types of genres, the student writing ability, the scale functioning, and rater performance might have been different. Although it is not clear to what extent these issues affected the current findings, these issues should be carefully taken into consideration when interpreting and generalising the current findings and when conducting scale development and validation research.

## CONCLUSION

This quantitative research employs many-facets Rasch measurement and argument-based validation frameworks with a view to building an initial validity argument for the newly developed binary and analytic scales designed specifically for formative classroom assessment in a Thai EFL university setting. The current MFRM analysis provides various useful indicators used as backing evidence for an initial validity argument for the rating scales. Firstly, the rating scales and raters generally provided accurate ratings, which supports the evaluation inference and responds to Research Questions 1 and 2, respectively. Secondly, the rating scales and raters largely generated consistent ratings, which contributes to the generalization inference and responds to Research Questions 3 and 4, respectively. Thirdly, the rating scale criteria sufficiently captured the defined writing ability, contributing to the explanation inference and responding to Research Question 5. Interestingly, the present findings illuminated that the binary scale generally showed more desirable MFRM statistics than the analytic scale with respect to accuracy, consistency, and construct coverability. This consolidates Lukácsi's (2021) and Park

and Yan's (2019) conclusion that a binary scale shows a greater potential to mitigate rater inconsistency and cognitive load. However, it is still early to draw such a conclusion since there are variant characteristics of binary scales that differ in terms of, for example, scoring format, scale length and score category which were perceived by raters in previous research (Khamboonruang, 2020; Kim, 2010) to differently influence rater cognition. It is also worth noting that although the MFRM results revealed desirable psychometric properties of the rating scales, the rating scales might not be user-friendly and practical for the current raters and future users in the context of use. As evidenced in previous research (Khamboonruang, 2020; Kim, 2010), while psychometric results largely confirmed the quality of a rating scale, raters expressed some concerns over the functioning and practicality of certain features of a rating scale. Therefore, psychometric quality may not fully guarantee the functionability and practicality of the rating scale.

## IMPLICATIONS

The current findings offer useful implications for EFL classroom writing instruction and assessment as well as rating scale construction and validation research. The current findings revealed that, in general, the binary scale appears to provide more consistent and accurate rating scores than the analytic scale. Therefore, it should offer more detailed diagnostic information and be more practical and supportive for formative classroom assessment, where teachers are typically bombarded with workload and students need immediate feedback during an ongoing classroom. Alternatively, the analytic scale may be used for summative assessment which focuses more on broader writing ability domains. The complementary use of both rating scales should plausibly provide a fuller understanding of students' writing performances and more thorough information to inform teaching and learning. It is yet still early to conclude based solely on the current MFRM evidence that the binary scale would function better than the analytic scale. A lot more studies are still needed to systematically investigate and compare the effectiveness and impact of varying design features of binary scales and other types of rating scales for different assessment purposes in different EFL contexts. This would shed more light on which type of rating scale is particularly suitable for a particular purpose and in a particular context.

The MFRM results revealed many interesting findings about the effectiveness of the rating score categories. That is, the designed three-score category for the analytic criteria did not really represent equal proportions along the latent writing sub-construct as is assumed by raw scores, implying that MFRM provides more accurate estimates of language ability than raw score-based methods. Moreover, the score categories that did not well represent a unique quality level of writing sub-constructs need to be revised to optimise its effectiveness for future use in the classroom context. After operationalising the rating scales in the target context of use, future research should seek full-scale evidential backing for other types of inferences in order to examine a complete validity argument for the rating scales particularly when they are applied across different genres, tasks, students, and raters. In conjunction with MFRM and/or other psychometric methods, mixed-methods research should employ traditional qualitative methods (e.g., interview, stimulated recall, and/or think-aloud) and innovative eye-tracking technology to probe into raters' cognitive process in order to investigate whether raters apply

rating scales in an intended way. Feedback from the scale users' interview could provide further insight into the functioning, practicality, and usefulness of the rating scales apart from psychometric evidence and illuminate any problematic scale features that should be refined. As Kane (2013) pointed out, validation is an ongoing process and validity changes over time and is context-bounded. Scale developers and validators are encouraged to cyclically refine, operationalise, and validate a rating scale until achieving its optimal validity for a particular assessment context, and also draw on current validity and validation concepts to frame specific validation frameworks for various kinds of assessments in various contexts which would serve as a basis for scale development and validation research in wider contexts. In particular, the present research concludes that the interface of MFRM and argument-based validation frameworks offers a systematic and rigorous means of validating rater-mediated language assessment.

## ACKNOWLEDGEMENT

This research project was financially supported by Mahasarakham University Research Grant. I would also like to thank Mike Linacre for always and kindly answering my questions and giving me very helpful suggestions about many-facets Rasch measurement and FACETS.

## THE AUTHOR

***Apichat Khamboonruang*** (https://orcid.org/0000-0002-7182-3501) is a lecturer of English in the Department of Western Languages and Linguistics at Mahasarakham University. He holds a PhD in Applied Linguistics from the University of Melbourne. His research interests include rater-mediated language assessment and language test development and validation.
*apichat.k@msu.ac.th*

## REFERENCES

American Education Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54–74. https://doi.org/10.1080/15434300903464418

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice, 18*(3), 279–293. https://doi.org/10.1080/0969594X.2010.526585

Chapelle, C. A. (2021). Validity in language assessment. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of language testing* (pp. 11–20). Routledge.

Chapelle, C. A., & Voss, E. (2021). *Validity argument in language testing: Case studies of validation research*. Cambridge University Press.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behaviour. *Language Assessment Quarterly, 9*(3), 270–292. https://doi.org/10.1080/15434303.2011.649381

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang.

Eckes, T. (2019). Many-facet Rasch measurement: Implications for rater-mediated language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment volume I: Fundamental techniques* (pp. 152–175). Routledge.

Engelhard, J. G., & Wind, S. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5–29. https://doi.org/10.1177/0265532209359514

Ghalib, T. K., & Al-Hattami, A. A. (2015). Holistic versus analytic evaluation of EFL writing: A case study. *English Language Teaching, 8*(7), 225–236. http://dx.doi.org/10.5539/elt.v8n7p225

Han, T. (2017). Scores assigned by inexpert EFL raters to different quality EFL compositions, and the raters' decision-making behaviors. *International Journal of Progressive Education, 13*(1), 136–152.

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice, 20*(3), 281–307. https://doi.org/10.1080/0969594X.2012.742422

Isbell, D. R. (2017). Assessing C2 writing ability on the certificate of English language proficiency: Rater and examinee age effects. *Assessing Writing, 34*, 37–49. http://dx.doi.org/10.1016/j.asw.2017.08.004

Jeong, H. (2017). Narrative and expository genre effects on students, raters, and performance criteria. *Assessing Writing, 31*, 113–125. https://doi.org/10.1016/j.asw.2016.08.006

Jeong, H. (2019). Writing scale effects on raters: An exploratory study. *Language Testing in Asia, 9*(20), 1–19. https://doi.org/10.1186/s40468-019-0097-4

Jiuliang, L. (2014). Examining genre effects on test takers' summary writing performance. *Assessing Writing, 22*, 75–90. http://dx.doi.org/10.1016/j.asw.2014.08.003

Jönsson, A., Balan, A., & Hartell, E. (2021). Analytic or holistic? A study about how to increase the agreement in teachers' grading. *Assessment in Education: Principles, Policy & Practice, 28*(3), 212–227. https://doi.org/10.1080/0969594X.2021.1884041

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M. T. (2021). Articulating a validity argument. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed., pp. 32–47). Routledge.

Khamboonruang, A. (2020). *Development and validation of a diagnostic rating scale for formative assessment in a Thai EFL university writing classroom: A mixed methods study* [Doctoral dissertation, The University of Melbourne]. Minerva Access. http://hdl.handle.net/11343/252672

Kim, Y.-H. (2010). *An argument-based validity inquiry into the empirically-derived descriptor-based diagnostic (EDD) assessment in ESL academic writing* [Doctoral dissertation, The University of Toronto]. TSpace. https://hdl.handle.net/1807/24786

Knoch, U. (2016). Validation of writing assessment. In C. A. Chapelle (Ed.), *Encyclopedia of applied linguistics* (pp. 1–6). Blackwell Publishing Ltd. https://doi.org/10.1002/9781405198431.wbeal1480

Knoch, U. (2021). Assessing writing. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed., pp. 236–253). Routledge.

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

Knoch, U., Deygers, B., & Khamboonruang, A. (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. https://doi.org/10.1177/0265532221994052

Knoch, U., Fairbairn, J., & Jin, Y. (2021). *Scoring second language spoken and written performance: Issues, options, and directions*. Equinox.

Lamprianou, I., Tsagari, D., & Kyriakou, N. (2021). The longitudinal stability of rating characteristics in an EFL examination: Methodological and substantive considerations. *Language Testing, 38*(2), 273–301. https://doi.org/10.1177/0265532220940960

Li, W. (2022). Scoring rubric reliability and internal validity in rater-mediated EFL writing assessment: Insights from many-facet Rasch measurement. *Read and Writing*. https://doi.org/10.1007/s11145-022-10279-1

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

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 258–278). JAM Press.

Linacre, J. M. (2022). *Facets computer program for many-facet Rasch measurement, version 3.84.0.* Winsteps.com.

Lukácsi, Z. (2021). Developing a level-specific checklist for assessing EFL writing. *Language Testing, 38*(1), 86–105. https://doi.org/10.1177/0265532220916703

Mahshanian, A., Eslami, A., & Ketabi, S. (2017). Raters' fatigue and their comments during scoring writing essays: A case of Iranian EFL learners. *Indonesian Journal of Applied Linguistics, 7*(2), 302–314. https://doi.org/10.17509/ijal.v7i2.8347

Mendoza, A., & Knoch, U. (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. https://doi.org/10.1016/j.asw.2017.12.003

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. http://dx.doi.org/10.1037/0003-066X.50.9.741

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.

Park, H., & Yan, X. (2019). An investigation into rater performance with a holistic scale and a binary, analytic scale on an ESL writing placement test. *Papers in Language Testing and Assessment, 8*(2), 34–64.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. The Danish Institute of Educational Research.

Şahan, Ö., & Razı, S. (2020). Do experience and text quality matter for raters' decision-making behaviors? *Language Testing, 37*(3), 311–332. https://doi.org/10.1177/0265532219900228

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3–12. https://doi.org/10.1093/elt/49.1.3

Wagner, M. (2015). *The centrality of cognitively diagnostic assessment for advancing secondary school ESL students' writing: A mixed methods study* [Doctoral dissertation, The University of Toronto]. TSpace. https://hdl.handle.net/1807/69530

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

Wiseman, C. S. (2012). A comparison of the performance of analytic vs. holistic scoring rubrics to assess L2 writing. *Iranian Journal of Language Testing, 2*(1), 59–92.

Yan, X., & Chuang, P.-L. (2022). How do raters learn to rate? Many-facet Rasch modeling of rater performance over the course of a rater certification program. *Language Testing*. https://doi.org/10.1177/02655322221074913

Zhu, Y., Fung, A. S. L., & Yang, L. (2021). A methodologically improved study on raters' personality and rating severity in writing assessment. *SAGE Open*, 1–16. https://doi.org/10.1177/21582440211009476

## Appendix A

## The finalised binary rating scale

| Writing domains | No. | Descriptors | Rating | |
|---|---|---|---|---|
| Main idea | 01 | The main idea is clear. | Yes | No |
| | 02 | The main idea relates to the topic or responds to the prompt. | Yes | No |
| Supporting ideas | 03 | All supporting ideas are clear. | Yes | No |
| | 04 | All supporting ideas relate to the topic or respond to the prompt. | Yes | No |
| | 05 | All supporting ideas are convincing. | Yes | No |
| | 06 | The supporting ideas are sufficiently provided. | Yes | No |
| Specific details | 07 | The specific details for all supporting ideas are clear. | Yes | No |
| | 08 | The specific details for all supporting ideas relate to the supporting ideas. | Yes | No |
| | 09 | The specific details for all supporting ideas relate to the topic or respond to the prompt. | Yes | No |
| | 10 | The specific details for all supporting ideas are convincing. | Yes | No |
| | 11 | The specific details for all supporting ideas are sufficiently provided. | Yes | No |
| Concluding statement | 12 | The concluding statement clearly restates and paraphrases the main idea. | Yes | No |
| | 13 | The concluding statement clearly summarises and paraphrases the supporting ideas | Yes | No |
| | 14 | The concluding statement is concise. | Yes | No |
| Paragraph unity | 15 | The main idea, all supporting ideas, and all specific details relate to the topic or respond to the prompt. | Yes | No |
| | 16 | The main idea, all supporting ideas, and all specific details are arranged smoothly and logically. | Yes | No |
| Sentence use | 17 | A variety of simple, compound, and complex sentences are used. | Yes | No |
| | 18 | All sentences are clear, comprehensible, and meaningful in the context of use. | Yes | No |
| | 19 | All sentences do not have any grammatical and mechanical errors. | Yes | No |
| Vocabulary use | 20 | A wide range of words are observably used. | Yes | No |
| | 21 | Most words are clear, comprehensible, and meaningful in the context of use. | Yes | No |
| | 22 | Most words are appropriate for written language. | Yes | No |

## Appendix B

## The finalised analytic rating scale

| Writing domains | Rating score categories | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Main idea | The main idea may be clear but doesn't relate to the topic or respond to the prompt. | The main idea may relate to the topic or respond to the prompt but is not convincing. | The main idea is clear, relates to the topic or responds to the prompt, is convincing, and can be supported by supporting ideas and details. |
| Supporting ideas | Most supporting ideas are not clear and related to the single main idea or the prompt. | Most supporting ideas are clear, relate to the single main idea or respond to the prompt, but are not convincing, or not unique from one another, or not sufficiently provided. | All supporting ideas are clear, relate to the single main idea or responds to the prompt, are convincing and unique from one another, and are sufficiently provided. |
| Specific details | Most specific details for all supporting ideas are not clear and related to the supporting ideas, the single main idea, and the prompt. | Most specific details for all supporting ideas are clear and related to the supporting ideas, the single main idea, and the prompt, but are not convincing, or not unique from one another, or are not sufficiently provided. | The specific details for all supporting ideas are clear, relate to the supporting ideas, the single main idea, and the prompt, are convincing and unique from one another, and are sufficient and relatively well-balanced. |
| Concluding statement | The concluding statement may or may not restate the main idea and/or summarise the supporting ideas, but the main idea and supporting ideas are not sufficiently paraphrased. | The concluding statement restates the main idea and summarises the supporting ideas, but the main idea and supporting ideas are not sufficiently paraphrased, or the concluding statement doesn't end with relevant and reasonable thought. | The concluding statement is concise, clearly restates and paragraphs the main idea, clearly summarises and paraphrases the supporting ideas, and also ends with relevant and reasonable thought. |
| Paragraph unity | The main idea, all supporting ideas, and all specific details are not related or somewhat related to one another and to the prompt, but the ideas or sentences are not arranged smoothly and logically. | The main idea, all supporting ideas, and all specific details are mostly related to one another and to the prompt, are generally arranged smoothly and logically, but more transition words or phrases are needed to enhance the idea connection. | The main idea, all supporting ideas, and all specific details are related to one another and to the prompt, are arranged smoothly and logically, and are consistently linked by appropriate transition words or phrases. |
| Sentence use | The sentences may or may not include simple compound, and complex sentences, but most sentences are not clear, comprehensible, and meaningful in the context of use and also have many grammatical and mechanical errors. | The sentences include a variety of simple compound, and complex sentences, and most sentences are clear, comprehensible, and meaningful in the context of use, but may have few grammatical and mechanical errors. | A variety of simple, compound, and complex sentences are used and all sentences are clear, comprehensible, and meaningful in the context of use and do not have any grammatical and mechanical errors. |
| Vocabulary use | A limited range of words are used. Many words or expressions are not clear and meaningful in the context of use, and are not appropriate for written language. | A wide range of words are observably used but some words or expressions are not clear and meaningful in the context of use, and not all words or expressions are appropriate for written language. | A wide range of words are obviously used. All words or expressions are clear, comprehensible, and meaningful in the context of use and are appropriate for written language. A variety of word forms or expressions are observably used. |