LEARN Journal: Language Education and Acquisition Research Network
ISSN: 2630-0672 (Print) | ISSN: 2672-9431 (Online)
Volume: 16, No: 1, January – June 2023

Language Institute, Thammasat University
https://so04.tci-thaijo.org/index.php/LEARN/index

EFL Thai Students' Rating Performances in Self- and Peer-Assessments of Writing: A Many-Facets Rasch Analysis

Apichat Khamboonruang

apichat.k@msu.ac.th, Department of Western Languages and Linguistics, Faculty of Humanities and Social Sciences, Mahasarakham University, Thailand

APA Citation:
Khamboonruang, A. (2023). EFL Thai students' rating performances in self- and peer-assessments of writing: A many-facets Rasch analysis. *LEARN Journal: Language Education and Acquisition Research Network, 16*(1), 221-245.

Received
26/04/2022

Received in revised form 07/11/2022

Accepted
06/12/2022

ABSTRACT

Despite the value of self- and peer-assessments in supporting learning, research reports mixed findings about self- and peer-assessments of writing, raising intriguing queries about the validity and utility of student-led assessment in a writing classroom. Applying a many-facets Rasch measurement approach, this quantitative study investigated 53 EFL Thai undergraduates' rating performances in self- and peer-assessments of writing during an ongoing classroom in a university setting. Main findings revealed that self-assessors tended to rate more leniently and heterogeneously than peer-assessors whose severity levels were far more homogeneous and closer to the level of teacher severity. Across self- and peer-ratings, students could observably distinguish writing performance quality, thereby assigning varying scores across rubric criteria, and lower-ability students were more variable in their levels of severity than those of high ability. The findings highlight the main and interaction effects of student-led assessment type and writing ability level on students' rating performances. That is, students' rating performances vary according to different types of student-led assessment and different levels of ability. More specifically, assessment type exercises a greater bearing on lower-ability raters' rating variability, while ability level exerts a stronger influence on self-raters' rating variability. The findings provide implications for student-led assessment design and research in the classroom context.

Keywords: rating performances, self-assessment, peer-assessment, EFL writing assessment, many-facets rasch analysis

Introduction

With the increased call for assessment-integrated teaching to promote learning in an ongoing classroom, students should thus play an active role not only in learning but also in assessment during a classroom process (Andrade, 2019; Andrade & Heritage, 2018; Christison, 2018; Edwards, 2013; Harris & Brown, 2018; Oscarson, 2013; Topping, 2013). Particularly in higher education, students need to engage more actively in and take more responsibility for their learning (Harris & Brown, 2018). As such, they need to be inculcated with a sense of self-regulated and autonomous learning which is of help to any kinds of learning (Oscarson, 2013; Topping, 2013). Teachers are, therefore, encouraged to utilise alternative self-assessment (SA) and peer-assessment (PA) which allow students to become actively involved in learning and assessment and develop a sense of self-regulated and cooperative learning (Al-Mahrooqi & Denman, 2018; Oscarson, 2013; Topping, 2013). Despite having long been studied and used, SA and PA have gained even greater attention and an ever-growing research interest in the fields of language teaching and assessment (Andrade, 2019; Li & Zhang, 2021; Ross, 1998).

A body of research has thus far examined SA and PA performances and impacts on student learning and achievement (e.g., Andrade, 2019; Erman Aslanoglu et al., 2020; Esfandiari & Myford, 2013; Han & Riazi, 2018; Hung et al., 2016; Khamboonruang, 2020; Li & Zhang, 2021; Matsuno, 2009; Ross, 1998; Saito, 2008; Saito & Fujita, 2004). In general, much research has shown a positive association of SA and PA with learning and achievement (e.g., Andrade, 2019; Han & Riazi, 2018; Khamboonruang, 2020). On the one hand, such research has shown low reliability and validity of SA and PA, with self-assessors tending to demonstrate lower rating consistency, accuracy, and severity than peer- and teacher-assessors (e.g., Andrade, 2019; Li & Zhang, 2021). Intriguingly in writing assessment, empirical studies have revealed discordant findings regarding SA and PA performances (Erman Aslanoglu et al., 2020; Esfandiari & Myford, 2013; Matsuno, 2009; Saito & Fujita, 2004). Little is also known about SA and PA performances embedded within an ongoing classroom (e.g., Saito & Fujita, 2004) as well as the impact of proficiency level on SA and PA performances (e.g., Matsuno, 2009). Particularly in the Thai EFL context, studies on SA and PA are very rare in the international literature (e.g., Khamboonruang, 2020), let alone those applying a many-facets Rasch measurement (MFRM) psychometric, which offers more reliable, valid, and fair estimates than other statistical techniques based on raw scores.

Following a MFRM approach, this study quantitatively investigates the rating behaviours of 53 EFL Thai students in SA and PA in a university writing classroom. Specifically, this study looked at three commonly-studied rater effects, namely severity, inconsistency, and restriction of range, and compared such behaviours between SA and PA, between ability groups, and also with teacher-led assessment (TA). This research aimed to provide more insights into the main effects of writing ability level and student-led assessment type and their interaction effects on students' rating performances. The next section reviews the literature related to SA and PA with emphasis on research using MFRM to examine self- and peer-assessments of writing.

Literature Review

Self- and Peer-Assessments

Self-assessment (SA) has been defined in different ways. According to Brown and Harris (2013, p. 368), SA is a descriptive and evaluative act that students perform with regard to their own work and academic abilities. SA may be viewed as an internal approach in which learners self-assess their own performance or work and monitor their own progression or achievement which may be compared to external criteria such as other relevant language tests and teacher-led assessment (Oscarson, 2013). From the cognitive and constructivist perspectives, SA is viewed as an integral part of self-regulated learning in that it promotes conscious, motivated, and aware engagement in cognitive and metacognitive learning strategies as well as self-reflection on learning strengths and weaknesses (Andrade, 2019; Harris & Brown, 2018; Oscarson, 2013). SA may be implemented through varying forms and activities and may be applied as outcome-based summative assessment of learning (Midraj, 2018; Oscarson, 2013) or process-based formative assessment for learning (Andrade, 2019; Oscarson, 2013). In the latter regard, SA should be viewed as a crucial strategy within formative assessment that focuses on promoting self-regulated learning and learning progression (Andrade & Heritage, 2018; Harris & Brown, 2018).

Peer assessment (PA) has also been defined in differing ways. Topping (2013, p. 395) defined PA as an arrangement for classmates to evaluate the performances or learning outcomes of their equal-status peers. Driven by constructivist-related second language acquisition theories, Topping and others have argued that the interactive and collaborative nature of the PA process opens more doors for students to actively and collaboratively learn and gain knowledge (Edwards, 2013; Sun & Doman, 2018; Topping, 2013). Like SA, PA can be carried out through a variety of forms and activities (Edwards, 2013; Topping, 2013) and conducted as formative or summative assessment and as qualitative or quantitative assessment (Topping, 2013). PA can take place in class or online, individually or in face-to-face pairs or groups, and/or during or after class time (Edwards, 2013; Sun & Doman, 2018). The PA results can be used for feedback, grading or both (Edwards, 2013). Through the PA process, students make a collaborative and mutual effort to construct knowledge and in turn develop higher order thinking and reasoning processes (Cheng & Warren, 2005). As with SA, the nature of PA is particularly well-suited for formative, learning-oriented, and performance-based assessments (Midraj, 2018; Oscarson, 2013).

However, SA and PA are not without limitations. SA and PA may not provide optimal benefits when students are not always and actively engaged in SA and PA processes (Edwards, 2013; Harris & Brown, 2018). As with other rater-mediated assessments, rater errors also undermine SA and PA reliability, validity, and fairness (Andrade, 2019; Andrade & Heritage, 2018; Harris & Brown, 2018; Li & Zhang, 2021; Ross, 1998). Empirical and systematic review studies have reported that self-and peer-raters were not as reliable and accurate as teacher-raters (Esfandiari & Myford, 2013; Saito & Fujita, 2004). In particular, students tended to show more lenient and heterogenous severity levels in SA (Esfandiari & Myford, 2013; Khamboonruang, 2020; Matsuno, 2009; Saito & Fujita, 2004), but more severe and homogeneous severity levels in PA (Esfandiari & Myford, 2013; Matsuno, 2009; Saito & Fujita, 2004).

A number of factors have been hypothesised to contribute to SA and PA variability. A critical review by Andrade (2019) found that SA consistency varies according to gender, age, assessment purpose (summative or formative), task type (generic or specific), criteria specificity, criteria explicitness (concrete or abstract), and ability level. Most of these factors have recently been supported by Li and Zhang's (2020) meta-analytic study. Nationality or cultural values could also be one of the many possible reasons underlying the overestimation and underestimation of SA and PA (Esfandiari & Myford, 2013; Matsuno, 2009). Additionally, students' awareness of assessment purposes and insufficient training and monitoring could contribute to students' rating variability (Andrade, 2019; Andrade & Heritage, 2018). When students know that SA results are used for grading, they are inclined to overestimate their performances (Andrade, 2019; Andrade & Heritage, 2018). Since it is, by nature, difficult to enhance SA and PA quality which is influenced by a variety of factors, SA and PA should be implemented as formative assessment, with emphasis on assessment process rather than assessment product, in order to promote self-regulated learning and learning progression, and in turn improve learning achievement in the classroom context (Andrade, 2019; Andrade & Heritage, 2018)

Many-Facets Rasch Measurement

To date, classical test theory, generalizability theory, and many-facet Rasch measurement (MFRM) theory are the main psychometric methods that have typically been employed in previous research to investigate rater behaviours. Yet, estimates from classical test theory and generalizability theory are based on raw scores which are prone to measurement errors, especially rater effects (McNamara et al., 2019; Myford & Wolfe, 2003). Consequently, estimates of rater severity vary according to levels of examinee ability, criterion difficulty, and task difficulty in a particular assessment context (Kim & Wilson, 2009; McNamara et al., 2019). Building on Rasch measurement theory (Rasch, 1960), Linacre (1989) developed a MFRM model which is capable of calibrating raw scores into equal-interval measurement units (called logits or measures) adjusted for variations across multiple sources or facets of score variability, such as rater, rubric, examinee, and task (Eckes, 2015). For example, test-taker proficiency logits are adjusted for differences across the levels of rater severity, criterion difficulty, and task difficulty. Similarly, rater severity logits are corrected for differences across the levels of test-taker proficiency, criterion difficulty, and task difficulty (Eckes, 2015). In this way, MFRM-based estimates are more reliable, valid and bias-free than estimates generated from classical test theory and generalizability theory. Unlike classical test theory and generalizability theory which require normally-distributed and complete data for robust estimation, MFRM is capable of handling non-normal and/or incomplete (missing) data, in which all raters do not score all test-takers' performances as long as the ratings assigned by individual raters are sufficiently linked (Eckes, 2015). Additionally, MFRM provides detailed analysis of rater effects, particularly severity/leniency, inconsistency, and restriction of range which are typically present and investigated in rater-mediated assessment. Severity is a rater's tendency to consistently give lower scores to examinees than they should receive, whereas leniency refers to a rater's tendency to consistently provide higher scores to examinees than they should obtain (Myford & Wolfe, 2003). When a rater tends to assign ratings deviating from other raters' ratings, the rater is deemed as exhibiting inconsistent ratings (Myford & Wolfe, 2003).

Restriction of range refers to a rater's tendency to assign scores within only a certain portion of a rubric. In other words, the rater fails to apply the entire range of scoring categories on a rubric (Myford & Wolfe, 2003). Undoubtedly, a MFRM technique offers more accurate and fine-grained research findings about rater behaviours.

Research on Self- and Peer-Assessments of Writing

This section focuses on reviewing previous research which applied MFRM to examine and compare rating performances between SA, PA, and TA in the writing assessment context.

Saito and Fujita (2004) compared severity levels between teachers' ratings and business-major undergraduates' SA and PA, which aimed at supporting ongoing learning in a Japanese classroom context. In their study, teachers (*N* = 2), self-raters (*N* = 47), and peer-raters (*N* = 45) used a four-point, six-criteria analytic scale to rate assignment essays on two tasks. Each student rated his/her essay and three double-blinded peer essays on two assignments in class, while each teacher scored all students' essays at a convenient time. Only the scores on the second assignment were used for data analysis. MFRM results revealed that teacher-raters were the most severe and their ratings correlated strongly and significantly with peer-ratings. Self-raters showed the highest severity variability and a weak rating correlation with teacher- and peer-ratings and were more severe than peer-raters.

Matsuno (2009) also examined rating behaviours between teachers (*N* = 4) and students (*N* = 91) in a Japanese EFL writing classroom, in which students' self- and peer-ratings were partly used for grading purposes. In her study, self-, peer-, and teacher-assessors received a rater training session and used a six-point, 12-criteria analytic scale to score the same pro-and-con assignment essay. Each student rated one of his/her essays and five blinded peer essays, while each teacher rated differing numbers of essays at their convenience. MFRM results showed that self-raters, especially high-ability raters, underestimated their writing ability. Peer-raters were the most lenient, underestimated high-ability students, and overestimated low-ability students. However, peer-raters' severity variability was not significantly affected by proficiency level and they were internally consistent in their ratings and assigned fewer biased ratings than self- and teacher-raters.

More recently, Esfandiari and Myford (2013) compared severity differences between teacher-assessors (*N* = 6) and student self- and peer-assessors (*N* = 188) set up for research purposes in two Iranian universities. After receiving a rater training session, self-, peer-, and teacher-raters applied a six-point, 15-criteria analytic scale to score the same opinion exam essays. Each student first self-rated one of his/her essays and subsequently peer-rated one different blinded essay in class, while individual teachers rated all students' essays at their convenient time. MFRM findings indicated that self- and peer-raters showed more variations in their levels of severity, and teacher- and peer-assessors were not significantly different in the levels of severity and tended to assign far more severe ratings than self-raters.

Recently, Erman Aslanoglu et al. (2020) investigated the rating behaviours of Turkish university students (*N* = 58) in self- and peer-assessments of a research proposal. Each student used an analytic rubric to self-rate his/her proposal and peer-rate group-work proposals. MFRM results showed that self- and peer-raters differed significantly in the levels of severity, with students tending to be more lenient in self-assessment but more severe in peer-assessment. Furthermore, and their peer-ratings were more consistent than their self-ratings.

Observably, the four studies reported both similar and discordant findings. It is however difficult to compare such findings and draw logical conclusions about students' self- and peer-rating behaviours across the studies since TA, SA, and PA within and between the studies were implemented under rather different contexts (e.g., Japanese, Iranian, and Turkish) and varying assessment conditions (e.g., in class or at home scoring; same or different essays for scoring; single or repeated assessment; and ordering of self- and peer-ratings). For example, while Esfandiari and Myford (2013) and Matsuno (2009) investigated a one-off SA and PA session which was not part of ongoing learning in the classroom, Saito and Fujita (2004) examined repeated SA and PA in an ongoing classroom but then used only the second-assignment scores for their MFRM analysis. It is also not clear whether each student rated the same set of peer essays and whether self-rating or peer-rating was conducted first during the scoring session in the studies conducted by Erman Aslanoglu et al. (2020), Matsuno (2009), and Saito and Fujita (2004). The scoring order could have impacted students' self- and peer-ratings. Furthermore, not all the studies reported the use of rater training in TA, SA, and PA, and teacher-raters in certain studies were not classroom teachers. Moreover, only Matsuno's study reported that proficiency level affected self-rating variability, particularly for very high-achieving students, but did not influence peer-rating variability. It could be plausible that the high variability between TA, SA and PA conditions and practices within and across the studies could differentially have influenced raters' decision-making behaviours, potentially resulting in variations in rating scores used for the MFRM analysis and hence, mixed findings. As a result, it is difficult to generalise these findings to other assessment contexts, thereby necessitating further research looking at SA and PA in different local contexts.

Rationale for the Study

It may be argued that much empirical and systematic review research has well established that SA and PA tend to show overestimation and low consistency in comparison to external measures and thus this issue has probably been adequately researched (Andrade, 2019; Li & Zhang, 2021). However, empirical research on self- and peer-assessments of writing has showed discordant findings, which suggests varying context-related and student-internal factors underlying SA and PA variations (Erman Aslanoglu et al., 2020; Esfandiari & Myford, 2013; Matsuno, 2009; Saito & Fujita, 2004). In spite of a plethora of studies comparing rater behaviours between teacher-, self- and peer-assessors, far fewer have focused on writing assessment and deployed MFRM to examine rater behaviours and little is also known about the impact of proficiency on students' SA and PA ratings. There is, therefore, a need to employ MFRM to examine the influence of SA and PA, together with proficiency level, on students' rating performances in varying writing assessment contexts and conditions. In particular, little research has explored EFL Thai students' rating behaviours and no research has yet to apply MFRM to investigate EFL Thai learners' self- and peer-rating behaviours. All this calls for further research in this line with a view to achieving more empirical findings that would throw more and novel light on students' SA and PA performances and therefore inform the design of meaningful SA and PA which can be integrated with and supportive of regular classroom learning and teaching.

Building on this line of research, this study employed a MFRM approach to investigate EFL Thai undergraduates' SA and PA behaviours in a writing classroom. In this study, SA is defined as the use of a rubric to evaluate one's own writing performance in a low-stakes formative assessment context in order to encourage individual students to reflect and improve on writing and learning and promote individual learners' self-regulated learning during an ongoing course. PA is defined as the use of a rubric to evaluate peers' writing performances. This study, therefore, aimed to address three research questions: *(1) Did students' severity levels differ between self- and peer-assessments and ability groups?; (2) Were students' ratings consistent between self- and peer-assessments and ability groups?;* and *(3) Did students' ratings show restriction of range between self- and peer-assessments and ability groups?*

Methodology

Context and Participants

This study was situated in an EFL writing classroom in a Thai public research-based university in 2021. The participants consisted of 53 second-year ELT students enrolled in an English composition course from three intact classrooms taught by the author, who had a writing teaching experience of about three years. The students comprised 15 males and 38 females, all of whom had never taken an English writing course and done SA and PA before. The students were also given a consent form to sign. The course was conducted online via Google Meet due to the COVID-19 pandemic. The students were taught sentence writing during the first half of the course and then paragraph writing during the second half.

Data Collection

The data were collected during the second half of the course where students were assigned to write a one-prompt opinion task (see Appendix A) first and subsequently a cause-effect task with two optional prompts (see Appendix B). Both tasks were scored using an analytic rubric developed based on teacher intuition and the course syllabus. The tasks and rubric were developed by the teacher. The rubric (see Appendix C) consisted of nine rating criteria, each of which was rated on a three-category scale: weak (1), improving (2), and satisfying (3).

At the beginning of the paragraph instruction, the students were informed that SA and PA were aimed at promoting their self-regulated and collaborative learning, that the SA and PA scores were not used as part of their grading, and that the teacher would use the same rubric when evaluating the students' works. The students were also trained on how to interpret and apply the rubric and the teacher showed them how to rate one example paragraph. However, the students were not necessarily expected to clearly and congruently understand the rating criteria during the training as it was assumed that they would, while learning over the course, gradually better comprehend the criteria, which were part of the learning contents. In fact, it was difficult to fully standardise and monitor self-and peer-rating procedures to ensure SA and PA quality since the study was situated within an ongoing writing classroom and only one teacher was responsible for teaching 53 students from three classrooms.

During the paragraph instruction, the students were assigned to write two drafts for each task. Each student was encouraged to use the rubric as a guideline for writing, self-evaluating, and revising the first draft before submitting it, together with its self-rated rubric. The teacher used the rubric to rate and comment on all students' first drafts and then retuned the drafts to the students, who then produced improved second drafts through the same SA process. The teacher used the rubric to score all students' second drafts and the scores were used for students' achievement grading. Due to the difficulty in doing PA online, the PA, initially aimed at promoting the students' collaborative learning, was instead conducted at the end of the course only for the current research purpose of comparing the rating performances between the self-assessors, peer-assessors, and teacher-assessor. All the students joined the same online PA session to score two peers' blinded second-draft paragraphs on two tasks. Although each student did not rate all peers' paragraphs, the peer-rating was designed to link all types of raters' ratings in order to meet the MFRM requirement for statistical estimation. Since the students were able to peer-rate only the second draft, and were under time constraints, only the second-draft scores from both tasks were used in this study, including the scores from the formative use of SA and the scores from the one-off PA session.

To gain more insight into the students' rating behaviours, the students were classified into three writing ability groups based on the teacher's second-draft writing scores combined from both tasks (see Table 1). The full score on each task was 27, thus making a total score of 54. However, the maximum score obtained by this group of students was 44 while the minimum was 29. Based on this range, students receiving scores from 40 to 44 were put into a high-ability group, those having scores within the range of 35-39 were placed into a moderate-ability group, and those obtaining scores from 29 to 34 were put into a low-ability group. Accordingly, there were 10, 27, and 16 students in the high-, mid-, and low-ability groups, respectively.

Table 1

Characteristics of High-, Mid-, and Low-Ability Groups

Ability group	Score range	Number of students
High (1 male, 9 females)	40-44	10
Mid (8 males, 19 females)	35-39	27
Low (6 males, 10 females)	29-34	16
All (15 males, 38 females)	19-45	53

The three ability groups and two assessment types resulted in seven types of raters altogether: teacher-assessment (TA), high-ability self-assessment (HSA), mid-ability self-assessment (MSA), low-ability self-assessment (LSA), high-ability peer-assessment (HPA), mid-ability peer-assessment (MPA), low-ability peer-assessment (LPA). Table 2 summarises the characteristics of the ratings assigned by seven types of raters.

Table 2

Ratings Assigned by Seven Types of Raters

Types of raters	Total rated students	Number of paragraphs	Total rated paragraphs	Total rated criteria	Total ratings
TA (*n* = 1)	53	2	106	9	954
HSA (*n* = 10)	10	2	20	9	180
MSA (*n* = 27)	27	2	54	9	486
LSA (*n* = 16)	16	2	32	9	288
HPA (*n* = 10)	2	4	40	9	360
MPA (*n* = 27)	2	4	108	9	972
LPAL (*n* = 16)	2	4	64	9	576

Data Analysis

The current MFRM analysis was run in the Facets programme (Version 3.83.6; Linacre, 2021) and was based on a four-facet rating scale model which assumes that the structure of the three-point rating scale was the same for all rating criteria. Facets uses the maximum likelihood method for parameter estimation (Linacre, 2021). The main facets under analysis included (1) rater, (2) student, (3) task and (4) criteria. The rater facet of interest included seven types of raters which were further trichotomised into TA, PA, and SA rater types in a newly-specified dummy rater facet. Although the student facet, the object of measurement in typical assessments, is typically chosen to float on the logit scale, in this study the rater facet was instead allowed to float on the logit scale since this study focused on examining rater behaviours. The present MFRM analysis followed two main stages. First, Rasch assumptions were checked to ensure meaningful MFRM results. Second, graphical and numerical MFRM results were examined in order to probe into the rating behaviours of different types of raters.

Results

The MFRM results were grouped into three parts. Firstly, evidence of data-model fit and local independence Rasch assumptions was presented to ensure reliable MFRM results. Secondly, rater severity results were presented, followed by results pertaining to rater inconsistency and restriction of range.

Data-Model Fit and Local Independence

First of all, the unexpected standardised residuals and rater agreement were inspected to investigate data-model fit and local independence assumptions for the purpose of ensuring meaningful MFRM statistics. Of 3,816 valid ratings, 162 (4%) were associated with unexpected standardised residuals outside ±2 and 6 (0.1%) were associated with unexpected standardised residuals outside ±3, both below the expected maximum of 5% and 1% respectively to ascertain satisfactory global data-model fit (Linacre, 2021). The satisfactory data-model fit means that the ratings assigned by the raters generally matched the ratings expected by the Rasch model. In other words, there was only a small number of inappropriate ratings deviating from the Rasch expectations. Amongst the 34,380 interrater agreement opportunities, 16,135 (46.9%) represented the observed interrater agreement and 17,140.9 (49.9%) made up the expected interrater agreement. The fact that the observed agreement percentage was slightly below the expected agreement percentage suggests that different types of raters exhibited somewhat but not overly interdependent ratings; that is, they were independent of one another in assigning ratings (Eckes, 2015). Overall, the results suggested satisfactory global data-model fit and independent rating in the current data, thereby ensuring reliable MFRM statistics in this study (Fan & Bond, 2019).

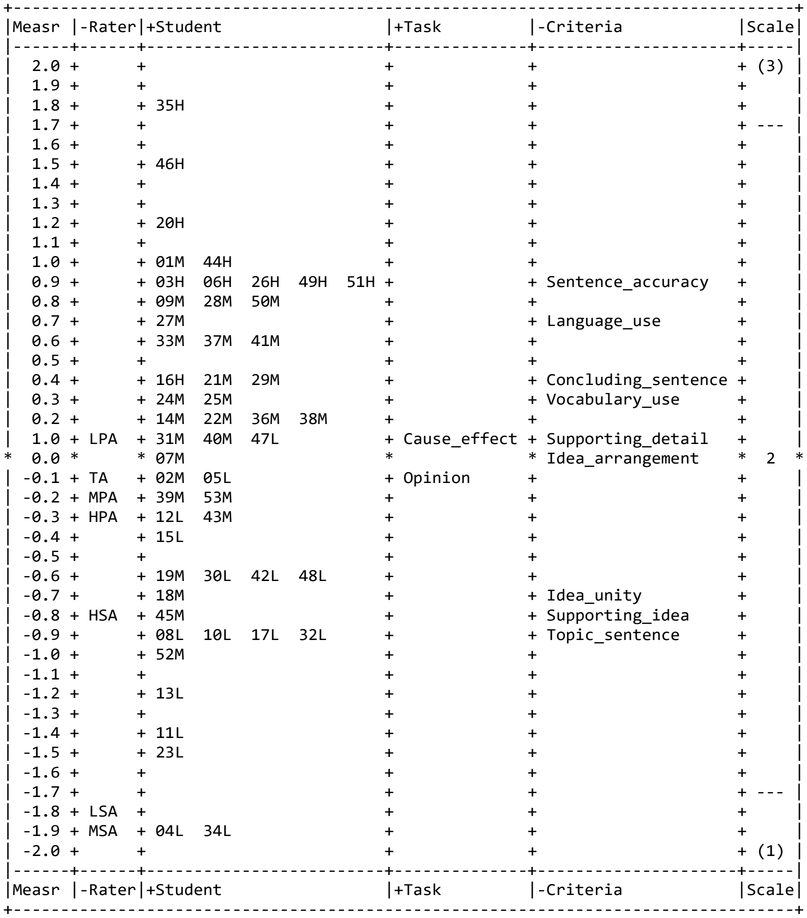
Rater Severity or Leniency

A variable map, fixed chi-square test, separation ratio, separation strata, and separation reliability were used to determine whether different types of raters differed significantly in the average levels of severity, contributing to Research Question 1.

Figure 1 displays the variable map which shows the levels of rater severity, student ability, task difficulty, and criterion difficulty, calibrated on the common standardised measure (henceforth referred to as logit) scale in the first column. The logit scale centres at 0 and ranges between 2 and -2 logits in this map. Higher logits represent higher levels of rater severity, student ability, task difficulty, and criterion difficulty. Overall, the map shows a wide spread of rater severity, student ability, and criterion difficulty on the logit scale. The three score categories (1, 2, and 3) in the last column were scaled in a desired order of difficulty; that is, higher scores, which are harder and thus require higher ability to achieve, were positioned above lower scores (Linacre, 2021). The students were labelled as low (L), mid (M), and high (H) ability based on the grouping criteria presented in Table 1. In general, student ability levels based on logits, adjusted for severity variations, were consistent with those derived from raw scores although some were inconsistent (e.g., 16H, 47L, 52M). This implies that ability estimates based on raw scores may not accurately capture and differentiate students' ability.

Figure 1

Variable Map Displaying Locations of Elements Within Each Facet



```
+----------------------------------------------------------------------------------+
|Measr |-Rater|+Student                 |+Task         |-Criteria            |Scale|
|------+------+-------------------------+--------------+---------------------+-----|
|  2.0 +      +                         +              +                     + (3) |
|  1.9 +      +                         +              +                     +     |
|  1.8 +      + 35H                     +              +                     +     |
|  1.7 +      +                         +              +                     + --- |
|  1.6 +      +                         +              +                     +     |
|  1.5 +      + 46H                     +              +                     +     |
|  1.4 +      +                         +              +                     +     |
|  1.3 +      +                         +              +                     +     |
|  1.2 +      + 20H                     +              +                     +     |
|  1.1 +      +                         +              +                     +     |
|  1.0 +      + 01M  44H                +              +                     +     |
|  0.9 +      + 03H  06H  26H  49H  51H +              + Sentence_accuracy   +     |
|  0.8 +      + 09M  28M  50M           +              +                     +     |
|  0.7 +      + 27M                     +              + Language_use        +     |
|  0.6 +      + 33M  37M  41M           +              +                     +     |
|  0.5 +      +                         +              +                     +     |
|  0.4 +      + 16H  21M  29M           +              + Concluding_sentence +     |
|  0.3 +      + 24M  25M                +              + Vocabulary_use      +     |
|  0.2 +      + 14M  22M  36M  38M      +              +                     +     |
|  1.0 + LPA  + 31M  40M  47L           + Cause_effect + Supporting_detail   +     |
*  0.0 *      * 07M                     *              * Idea_arrangement    *  2  *
| -0.1 + TA   + 02M  05L                + Opinion      +                     +     |
| -0.2 + MPA  + 39M  53M                +              +                     +     |
| -0.3 + HPA  + 12L  43M                +              +                     +     |
| -0.4 +      + 15L                     +              +                     +     |
| -0.5 +      +                         +              +                     +     |
| -0.6 +      + 19M  30L  42L  48L      +              +                     +     |
| -0.7 +      + 18M                     +              + Idea_unity          +     |
| -0.8 + HSA  + 45M                     +              + Supporting_idea     +     |
| -0.9 +      + 08L  10L  17L  32L      +              + Topic_sentence      +     |
| -1.0 +      + 52M                     +              +                     +     |
| -1.1 +      +                         +              +                     +     |
| -1.2 +      + 13L                     +              +                     +     |
| -1.3 +      +                         +              +                     +     |
| -1.4 +      + 11L                     +              +                     +     |
| -1.5 +      + 23L                     +              +                     +     |
| -1.6 +      +                         +              +                     +     |
| -1.7 +      +                         +              +                     + --- |
| -1.8 + LSA  +                         +              +                     +     |
| -1.9 + MSA  + 04L  34L                +              +                     +     |
| -2.0 +      +                         +              +                     + (1) |
|------+------+-------------------------+--------------+---------------------+-----|
|Measr |-Rater|+Student                 |+Task         |-Criteria            |Scale|
+----------------------------------------------------------------------------------+
```

Table 3 shows the severity heterogeneity statistics of the seven rater types. As can be seen, the significant fixed chi-square test ($X^2(6)$ 548.4, $p <$ 0.01) indicates that at least two types of raters differed significantly in severity (Eckes, 2015). The very high rater separation ratio (8.82) and strata (12.09) values, far greater than the expected value of 1, suggest that the raters' severity levels could be stratified into about 12 statistically distinct classes. The very high separation reliability (0.99) suggests that the rater separation indices were highly reliable (Eckes, 2015). Furthermore, there was a significant variability in the levels of severity within the SA raters, as indicated by the significant fixed chi-squared test ($X^2(2)$ 45.1, $p < 0.01$) as well as the rather high rater separation ratio (5.19), strata (7.25), and reliability (0.96). A noticeable variability in the levels of severity was also found within the PA raters, as suggested by the significant fixed chi-squared test ($X^2(2)$ 12.8, $p < 0.01$), along with the relatively high rater separation ratio (2.51), strata (3.67), and reliability (0.86). However, the severity variations within the PA raters were not as strong as those within the SA raters, which were almost double those exhibited by the PA raters.

Table 3

Severity Heterogeneity Indicators

Severity calibration	All rater types	SA raters	PA raters
Separation ratio	8.82	5.19	2.51
Separation strata	12.09	7.25	3.67
Separation reliability	0.99	0.96	0.86
Fixed chi-square test	548.4, *df* = 6 *p* = 0.00	45.1, *df* = 2, *p* = 0.00	12.8, *df* = 2, *p* = 0.00

Table 4 lays out the rater statistics arranged in descending order of severity level. The standard errors of estimates (SE) were very close to 0, supporting a rather precise estimation of the severity logits (Eckes, 2015). Amongst the seven rater types, the LPA raters, showing the highest logit of 0.09, were the most severe (or least lenient), whereas the MSA raters, showing the lowest logit of -1.94, were the least severe (or most lenient). Interestingly, the TA and PA raters' severity logits were closely clustered, suggesting their rather homogenous ratings.

Table 4

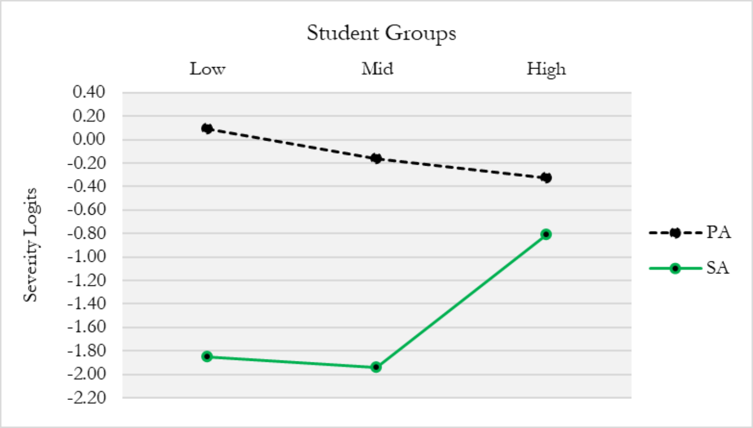
Rating Counts and Severity Estimates

		Severity estimates	
Rater types	Rating counts	Logit	*SE*
LPA	576	0.09	0.08
TA	954	-0.06	0.06
MPA	972	-0.16	0.06
HPA	360	-0.33	0.10
HSA	180	-0.81	0.15
LSA	288	-1.85	0.11
MSA	486	-1.94	0.09
TA	954	-0.06	0.06
PA	1908	-0.13	0.08
SA	954	-1.53	0.12

Figure 2 shows that students, irrespective of ability levels, displayed higher severity in PA than in SA, with more proficient raters showing more invariant severity levels than less proficient raters across SA and PA. More interestingly, low- and high-ability raters showed opposing patterns of severity levels. That is, while low-ability assessors scored most leniently in SA but most harshly in PA, high-ability raters rated most leniently in PA but most harshly in SA.

Figure 2

Rater Severity Logits Across SA , PA, and Ability Groups



Rater Inconsistency and Restriction of Range

A weighted mean square residual fit (Infit MnSq) statistic, which has a range from 0 to infinite, was employed to determine whether the raters were consistent and showed restriction of range in their ratings, responding to Research Questions 2 and 3, respectively.

As shown in Table 5, all rater types showed Infit indices close to the expected value of 1 and within the acceptable range of 0.50 - 1.50 (Linacre, 2021), suggesting appropriate ratings as expected by the Rasch model (Linacre, 2021). None of the raters showed an Infit value over 1.50, indicating acceptable consistency in their ratings (Myford & Wolfe, 2003). No rater types had an Infit value lower than 0.05, suggesting no significant restriction of the range in their ratings. In other words, all types of raters could acceptably differentiate the quality levels of writing performances and rating criteria and thus apply all rating points on the rubric (Myford & Wolfe, 2003). On a closer analysis, the LPA raters showed the highest Infit index of 1.43, meaning that their ratings were the least consistent, while the HSA assessors had the lowest Infit index at 0.72, implying that their ratings showed the highest, but still acceptable restriction of range (Linacre, 2021). The Infit index of the PA raters was 1.21, slightly over 1, suggesting more variations in peer-ratings than self- and teacher-ratings, which demonstrated noticeable yet acceptable restriction of range, as suggested by the Infit values of 0.78 and 0.81, respectively. As displayed in Figure 3, the PA raters showed higher Infit values than self-raters across ability groups, implying that the PA ratings were more varied than the SA ratings.

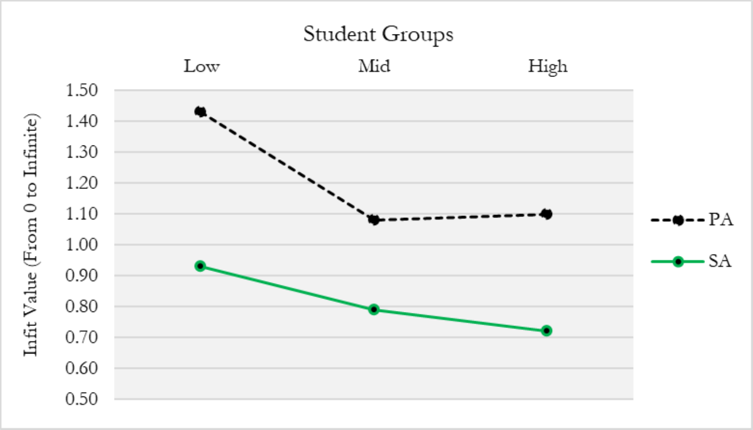
Table 5

Rater Infit Statistic

Rater types	Infit statistic
TA	0.78
MPA	1.08
HPA	1.10
MPA	1.08
LPA	1.43
HSA	0.72
MSA	0.79
LSA	0.93
TA	0.78
PA	1.21
SA	0.81

Figure 3

Rater Infit Indices Across SA , PA, and Ability Groups



Discussion and Conclusions

This study was probably the first applying many-facet Rasch measurement to investigate Thai EFL students' self- and peer-ratings in a higher education writing classroom, thereby providing novel and detailed findings about Thai EFL students' rating behaviours. This study specifically examined rater severity, inconsistency, and restriction of range effects within and between rater types and writing ability groups. The present study uncovered a number of interesting findings which both support and contradict those found in previous research.

Regarding rater severity, one of the present findings was that student self- and peer-ratings were generally more lenient than teacher-ratings, which mirrors similar findings by Esfandiari and Myford (2013) and Saito and Fujita (2004). However, this finding contradicts some previous findings that self-raters were more severe than peer-raters (Erman Aslanoglu et al., 2020; Matsuno, 2009; Saito & Fujita, 2004), and that peer-assessors tended to rate more leniently than self- and teacher-raters (Matsuno, 2009). The current results also support those of Matsuno (2009) and Saito and Fujita (2004) who, likewise, discovered that self-rater severity was more heterogenous than peer-rater severity. Moreover, the present study found that self-raters were more lenient than peer- and teacher-raters, which both confirms Esfandiari and Myford's (2013) study and contradicts the studies by Erman Aslanoglu et al. (2020), Matsuno (2009), and Saito and Fujita (2004) who found that peer-raters were more lenient than self-raters. Furthermore, the present findings revealed that peer-rater severity and teacher-rater severity were rather congruent, corresponding with the findings of Esfandiari and Myford (2013), Matsuno (2009), and Saito and Fujita (2004). In line with Matsuno's (2009) findings, this study revealed that self-raters showed a greater variability of severity than peer-raters. Intriguingly, low-ability raters generally showed an extremely lenient approach towards self-assessment but an extremely severe attitude towards peer-assessment, whereas high-ability raters demonstrated a reverse pattern of severity. Across assessment types and ability groups, students' ratings were not unduly redundant or inconsistent. That is, they were able to differentiate writing performance quality and rating criteria and thus did not overuse a limited range of the scoring points on the rubric.

A number of hypotheses have been made in previous research regarding factors underlying rating variability in students' self-and peer-assessments. For example, Matsuno (2009) hypothesised that students' high self-rating severity was influenced by their cultural value of modesty in Japan. Esfandiari and Myford (2013) argued, however, that such modesty is not valued in the Iranian context and thus did not influence students' overrating or underrating tendency; they therefore hypothesised that this tendency may instead have been derived from the norm-referenced course evaluation method. In the present srtudy, however, neither cultural values nor course evaluation method should have significantly influenced students' overestimation of self-assessment and underestimation of peer-assessment since students were informed that their self-rating and peer-rating scores were not used for course evaluation and that nobody knew each other's self- and peer-assessment scores. They also did not know who wrote the paragraphs they peer-rated. Instead, the rating condition situated within the self- and peer-assessments is one of the most likely sources underlying students' rating variability. As mentioned earlier, the students were allowed to self-score their works outside class and thus they rated their works under different self-rating conditions. Their effort and attention put into the self-assessment could also have varied. As for peer-assessment, the students rated their peers' work in the same single session under highly similar rating conditions. This rating condition could potentially have propelled the students to pay more attention and effort to peer-assessment than they paid to self-assessment and thus their peer-ratings became more homogeneous than their self-ratings. Due to this, it is impossible in the current assessment to disentangle rating condition from rater type since these factors could have both contributed to differences in the rating behaviours between self and peer-assessors. The rating ordering and ongoing learning might also have influenced the students' self- and peer-ratings. Before the peer-rating at the end of the course, the students had learned more contents, repeatedly self-rated their own paragraphs, and received feedback from the teacher, which could subsequently have contributed to their better performances of peer assessment in general. Indeed, it is difficult to accurately specify factors underlying variations in self- and peer-ratings in the present study, where student rating behaviours could be influenced by a variety of factors and variables. Apart from the student-led assessment type and ability level under study as well as the factors discussed above, there might be other factors or variables, both student-internal and student-external, that could have influenced variability in students' rating performances. However, this is beyond the scope of this study.

To conclude, the findings from this study pinpoint the main and interaction effects of student-led assessment type and writing ability level on students' rating bahviours; that is, student rating behaviour varies depending on different types of student-led assessment and different levels of writing ability. Specifically, ability level exerts a greater influence on self-assessment variability than peer-assessment variability, while assessment type exercises a greater bearing on low-ability raters. Self-raters are more lenient than peer- and teacher-assessors. Higher-ability students tend to be more severe in self-assessment but more lenient in peer-assessment vis-à-vis lower-ability students. All in all, peer-assessment is thus more reliable and accurate than self-assessment and more proficient raters are more reliable and accurate than less proficient raters.

Limitations

Before drawing conclusions and generalisations from the current findings, it is important to acknowledge certain caveats in this study. First of all, since this study was situated in a real-world classroom context where the number of students was small, the current dataset was also rather small, which might have influenced the current MFRM statistics. As McNamara et al., (2019) caution, an insufficient dataset undermines the precision of estimates and the robustness of fit statistics (pp. 146–147). In fact, there are no clear-cut rules of thumbs for sufficient sample size in a MFRM analysis. Although Barkaoui (2014, p. 15) suggests that a reasonable MFRM analysis should have 30 test-takers, 10 raters, and three tasks and each facet must contain at least two elements, he acknowledges that much existing research has based its MFRM analysis on much smaller datasets. The rather small dataset in this study notwithstanding, the Facets programme did not show any warnings or problems in the MFRM analysis. A second caveat would be that there was only one teacher who rated students' writing performances. If more teachers had scored students' writing performances, the average level of teacher severity might have been different. Thirdly, the criteria for grouping students' ability levels were based on the teacher' ratings, which were assumed to be reliable and valid in the classroom assessment context. If the rater grouping had been based on other criteria, the findings might have turned out differently. Fourthly, the findings might have been different if different rating designs had been used in the self-and peer-assessments and the students' peer-assessment had been carried out formatively during the course. Finally, if the students had had more opportunity to practice rating example paragraphs, they might have rated more reliably and accurately.

Implications

The current findings suggest that considering the highly-varied and non-standardised nature of classroom assessment, students' rating performances, particularly those of peer and high-ability raters, are reliable, albeit to a limited extent. Therefore, self- and peer-assessments can be considered appropriate assessment activities. However, the use of self-assessment, particularly amongst low-ability students, needs to be carefully considered, especially in a formal high-stakes assessment context. It is particularly suggested that peer-assessment and high-ability students' ratings should be considered more reliable and accurate than self-assessment and low-ability students' ratings when ascertaining meaningful assessment information and making important decisions. When both self- and peer-assessments are expected to inform high-stakes decisions about learning and teaching, teachers should involve students in rubric development and need to provide substantial rater training, which should include scoring practice, close monitoring, and detailed feedback over time, in order to help students rate more appropriately. Substantial rater training and close monitoring, however, may not be feasible in practice considering the real-world classroom context. Instead, teachers are highly encouraged to use student-led assessment for low-stakes formative assessment purposes with emphasis on self- and peer-assessment processes or activities rather than rating scores or outcomes in order to promote autonomous and colloborative learning. To ensure the optimal value of student-led formative assessment, teachers need to make sure that students are actively engaged in the self- and peer-assesssment processes.

Future research should investigate self- and peer-assessments fully integrated in formative assessment to gain a fuller understanding of students' rating behaviours and their potential in promoting self-regulated learning and achievement. More research is still needed to investigate EFL Thai learners' self- and peer-rating performances in order to gain a more profound insight into students' rating behaviours. Researchers are encouraged to experiment with ways to enhance students' rating performances and engagement in self- and peer-assessments. If possible, researchers should compare students' self- and peer-rating performances under similar rating conditions in order to determine to what extent rating condition might play a role in students' self- and peer-assessment behaviours. Researchers should employ qualitative methods (e.g., think-aloud and interview), together with data science technology (e.g., eye-tracking methods) to probe into factors underlying students' self- and peer-rating behaviours and how self- and peer-rating behaviours relate to improvement in student learning and achievement. It is also worthwhile to inspect whether different types of student self- and peer-raters exhibit psychometric evidence of bias, interaction, or differential rater functioning effects, which will shed more light on the validity and fairness of student-led assessment. Perhaps most fruitfully, researchers and educators are highly encouraged to incorporate advanced technology and psychometrics in designing, implementing, and researching innovative student-led assessment that would not only optimise students' rating performance and learning but also produce trustworthy and insightful research findings.

Acknowledgements

This research project was financially supported by Mahasarakham University Research Grant. My special appreciation also goes to Mike Linacre for always and kindly answering my questions and giving me very helpful suggestions regarding many-facets Rasch measurement and Facets programme.

About the Author

Apichat Khamboonruang: A lecturer of English in the Department of Western Languages and Linguistics at Mahasarakham University. He holds an MA in English as an International Language from Chulalongkorn University and a PhD in Applied Linguistics from the University of Melbourne. His research interests include rater-mediated language assessment and language test development and validation.
(ORCID iD: 0000-0002-7182-3501)

References

Al-Mahrooqi, R., & Denman, C. (2018). Alternative assessment. In J. L. Liontas & M. DelliCarpini (Eds.), *The TESOL encyclopedia of English language teaching* (pp. 1–6). Wiley-Blackwell. https://doi.org/10.1002/9781118784235.eelt0325

Andrade, H. L. (2019). A critical review of research on student self-assessment. *Frontiers in Education, 4*(87), 1–13. https://doi.org/10.3389/feduc.2019.00087

Andrade, H. L., & Heritage, M. (2018). *Using formative assessment to enhance learning, achievement, and academic self-regulation*. Routledge.

Barkaoui, K. (2014). Multifaceted Rasch analysis for test evaluation. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1–22). John Wiley & Sons, Inc. doi: 10.1002/9781118411360.wbcla070

Brown, G. T. L., & Harris, L. R. (2013). Student self-assessment. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 367–393). SAGE Publications, Inc.

Cheng, W., & Warren, M. (2005). Peer assessment of language proficiency. *Language Testing, 22*(1), 93–121. https://doi.org/10.1191/0265532205lt298oa

Christison, M. (2018). Student involvement in assessment. In J. L. Liontas & M. DelliCarpini (Eds.), *The TESOL encyclopedia of English language* *teaching* (pp. 1–7). Wiley-Blackwell. https://doi.org/10.1002/9781118784235.eelt0357

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

Edwards, J. G. H. (2013). Peer assessment in the classroom. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 730–750). Wiley-Blackwell. https://doi.org/10.1002/9781118411360.wbcla002

Erman Aslanoglu, A., Karakaya, I., & Sata, M. (2020). Evaluation of university students' rating behaviors in self and peer-rating process via many facet Rasch model. *Eurasian Journal of Educational Research*, 89, 25–46.

Esfandiari, R., & Myford, C. M. (2013). Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays. *Assessing Writing, 18*(2), 111–131. https://doi.org/10.1016/j.asw.2012.12.002

Fan, J., & Bond, T. (2019). Applying Rasch measurement in language assessment: Unidimensionality and local independence. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment volume I: Fundamental techniques* (pp. 83–102). Routledge.

Han, C., & Riazi, M. (2018). The accuracy of student self-assessments of English-Chinese bidirectional interpretation: A longitudinal quantitative study. *Assessment & Evaluation in Higher Education, 43*(3), 386–398. https://doi.org/10.1080/02602938.2017.1353062

Harris, L. R., & Brown, G. T. L. (2018). *Using self-assessment to improve student learning.* Routledge.

Hung, Y.-j., Samuelson, B. L., & Chen, S.-c. (2016). Relationships between peer- and self-assessment and teacher assessment of young EFL learners' oral presentations. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 317–338). Springer, Cham. https://doi.org/10.1007/978-3-319-22422-0

Kim, S. C., & Wilson, M. (2009). A comparative analysis of the ratings in performance assessment using Generalizability theory and the many-facet Rasch model. *Journal of Applied Measurement, 10*(4), 408–423.

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

Li, M., & Zhang, X. (2021). A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing, 38*(2), 189–218. https://doi.org/10.1177/0265532220932481

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

Linacre, J. M. (2021). Facets computer program for many-facet Rasch measurement, version 3.83.6. Winsteps.com

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*(1), 075–100. https://doi.org/10.1177/0265532208097337

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice, and language assessment: The role of measurement.* Oxford University Press.

Midraj, J. (2018). Self-assessment. In J. L. Liontas & M. DelliCarpini (Eds.), *The TESOL encyclopedia of English language teaching* (pp. 1–7). Wiley-Blackwell. https://doi.org/10.1002/9781118784235.eelt0331

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.

Oscarson, M. (2013). Self-assessment in the classroom. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 712–729). Wiley-Blackwell. https://doi.org/10.1002/9781118411360.wbcla046

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

Ross, S. (1998). Self-assessment in second language testing: a meta-analysis and analysis of experiential factors. *Language Testing, 15*(1), 1–20. https://doi.org/10.1177/026553229801500101

Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing, 25*(4), 553–581. https://doi.org/10.1177/0265532208094276

Saito, H., & Fujita, T. (2004). Characteristics and user acceptance of peer-rating in EFL writing classrooms. *Language Teaching Research, 8*(1), 31–54. https://doi.org/10.1191/1362168804lr133oa

Sun, Y., & Doman, E. (2018). Peer assessment. In J. L. Liontas & M. DelliCarpini (Eds.), *The TESOL encyclopedia of English language teaching* (pp. 1–7). Wiley-Blackwell. https://doi.org/10.1002/9781118784235.eelt0330

Topping, K. J. (2013). Peers as a source of formative and summative assessment. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 395–412). SAGE Publications, Inc.

Appendix A
Opinion Paragraph Writing Task

Task 11: Opinion Paragraph Writing

Write an opinion paragraph of between 150 and 200words in response to the prompt below:

Do you agree or disagree with the following statement? "It is impossible for Thai learners to improve English in Thailand where English is not the first and formal language." State your opinion and give three convincing and sufficient reasons and details to support your opinion.

Your paragraph writing needs to include (1) the title of your writing, (2) the topic sentence or main idea, (3) three main supporting points, (4) specific supporting details for the three main supporting points, and (5) the concluding sentence restating the main idea and summarising the main supporting points:

Outline or Template

Title of paragraph writing	
Background / introductory sentence(s)	
Topic sentence (your opinion)	
Main Supporting point (reason) 1	
• Specific supporting details	
Main Supporting point (reason) 2	
• Specific supporting details	
Main Supporting point (reason) 3	
• Specific supporting details	
Concluding sentence	
• Restate the topic	
• Summarise the main points or reasons	

Appendix B
Cause-Effect Paragraph Writing Task

Task 2: Cause-Effect Paragraph Writing

Choose only one of the following prompts below and write a cause-effect paragraph of between 150 and 200 words in response to chosen prompt.

1) *What do you think are possible reasons (causes) why **some students cheat** in the exam or test?*
2) *What do you think are the positive effects of **self-assessment activities** on students' writing learning?*

Read the chosen prompt carefully and critically before you determine whether to use the focus-on-causes method or the focus-on-effects method for your paragraph. As you develop your paragraph, try to **reread the writing question** and **follow the example paragraphs** in the teaching materials. Your paragraph writing needs to include (1) the title of your writing, (2) the topic sentence or main idea, (3) three main supporting points, (4) specific supporting details for each main supporting point, and (5) the concluding sentence which may restate the main idea, summarise the main supporting points, and/or offer suggestions or predictions.

Outline or Template

Title of a paragraph	
Topic Sentence	
Main supporting point 1	
• Specific supporting detail 1	
• Specific supporting detail 2	
Main supporting point 2	
• Specific supporting detail 1	
• Specific supporting detail 2	
Main supporting point 3	
• Specific supporting detail 1	
• Specific supporting detail 2	
Concluding sentence	
• Restate the topic	
• Summarise the main points or reasons	

Appendix C

Paragraph Writing Analytic Rubric

Paragraph writing skills	🙄	🙂	😀
	1 (Weak)	2 (Improving)	3 (Satisfying)
1. The main idea or topic sentence	The main idea or topic sentence is not clear and/or relevant to the topic or prompt. (1)	The main idea or topic sentence may be relevant but not clear and/or specific to the topic or prompt and not well-structured. (2)	The main idea or topic sentence is clear, relevant, and specific to the topic or prompt and well-structured. (3)
2. The supporting idea	The supporting ideas are not generally clear and/or relevant to the main idea, topic, or prompt, and are not sufficient and convincing. (1)	The supporting ideas may generally be clear and/or relevant to the main idea, topic, or prompt, but are not sufficient, convincing and/or well-arranged. (2)	The supporting ideas are clear and/or relevant to the main idea, topic, or prompt, and sufficient, convincing, and well-arranged. (3)
3. The specific detail	The specific details are not generally clear and/or relevant to the supporting ideas, main idea, or topic and are not sufficient, convincing and/or well-arranged. (1)	The supporting ideas may generally be clear and/or relevant to the supporting ideas, main idea, or topic but are not sufficient, convincing and/or well-arranged. (2)	The supporting ideas are clear and/or relevant to the supporting ideas, main idea, or topic and also sufficient, convincing, and well-arranged. (3)
4. Idea unity	Ideas are partly relevant to the topic or prompt. (1)	Ideas are generally relevant to the topic or prompt but are somewhat redundant or not clear or logical. (2)	Ideas are relevant to the topic or prompt and are clear or logical. (3)
5. Idea arrangement	Within-paragraph transitions are rarely used or are not mostly used appropriately. (1)	Within-paragraph transitions are mostly used appropriately. (2)	Within-paragraph transitions are all used appropriately. (3)
6. The concluding sentence	The concluding sentence restates the main idea or summarises the supporting points but is not well-paraphrased. (1)	The concluding sentence restates the main idea, and summarises the supporting points but is not well-paraphrased. (2)	The concluding sentence restates the main idea, summarises the supporting points, and is well-paraphrased. (3)
7. Sentence	Many sentences are not built accurately. (1)	Almost all sentences are built accurately. (2)	All sentences are built accurately. (3)
8. Vocabulary	A limited range of words are used and many words are not used appropriately. (1)	A wide range of words are used but some are not used appropriately. (2)	A wide range of words are appropriately used. (3)
9. Overall language	Overall language is generally clear but still has too many linguistic/grammatical problems. (1)	Overall language is almost all clear with certain linguistic/grammatical problems. (2)	Overall language is all clear without linguistic/grammatical problems. (3)