PASAA Journal
Volume 66, October 2023, 5-36

PASAA

Detecting Differential Rater Severity in a High-Stakes EFL Classroom Writing Assessment: A Many-Facets Rasch Measurement Approach

Apichat Khamboonruang

Chulalongkorn University Language Institute, Thailand
apichat.kh@chula.ac.th

Article information	
Abstract	Differential rater severity (DRS), one prevalent case of differential rater functioning (aka rater bias or rater interaction) effects, manifests itself when a rater assigns unusually severe or lenient ratings, threatening the validity and fairness of rater-mediated assessment. Building on a many-facets Rasch measurement (MFRM) approach, this study aimed to detect whether teachers exercised DRS towards rating criteria and student subgroups (classroom, proficiency, and gender) in a high-stakes EFL classroom writing examination. Data were collected from three teachers who applied a four-point five-criteria analytic rubric to rate opinion essays written by 42 English-major undergraduates during the examination. Main findings revealed that the teachers were not uniform in their severity levels, with the most experienced teacher exhibiting the highest severity and the least experienced teacher exercising the lowest severity. Whilst the most experienced and most severe teacher exposed slight DRS towards student genders, the less experienced and less severe teachers exerted substantial DRS in reverse pattern towards rating criteria and

E-ISSN: 2287-0024

	student classrooms. Surprisingly, the less experienced teachers scored their own classroom students less severely but marked each other's classroom students more severely than expected. The current findings raise the attention and awareness of teachers, educators, and policymakers concerning the impact of rater effects on the validity and fairness of rater-mediated assessment in the classroom context.
Keywords	differential rater functioning, differential rater severity, EFL classroom writing assessment, many-facets Rasch measurement
APA citation:	Khamboonruang, A. (2023). Detecting differential rater severity in a high-stakes EFL classroom writing assessment: A many-facets Rasch measurement approach [Special Issue]. *PASAA, 66,* 5–36.

1. Introduction

Differential rater functioning, also known as rater bias or rater interaction, is another form of rater effects in which a rater's judgemental tendency varies when the rater interacts with undesirable construct-irrelevant sources (e.g., examinee gender, examinee proficiency, and examinee race), threatening the validity and fairness of rater-mediated assessment (Engelhard & Wind, 2018). A prevalent case of differential rater functioning is differential rater severity or leniency, henceforth referred to as differential rater severity (DRS), where a rater rates particular subgroups of examinees systematically more or less severely than the rater usually does (Eckes, 2019; Myford & Wolfe, 2003). Research has revealed that raters were prone to exhibit not only different levels of severity (Khamboonruang, 2020; Youn, 2018) but also different patterns of DRS towards rating criteria (Eckes, 2005; Kondo-Brown, 2002; Schaefer, 2008; Wind & Engelhard, 2013) and certain subgroups of examinees (Eckes, 2005; Engelhard & Myford, 2003; Erman Aslanoğlu & Şata, 2021; Kondo-Brown, 2002; Wind & Sebok-Syer, 2019; Youn, 2018). In fact, severity and DRS effects result in inaccurate and unfair estimates of examinee performance or ability, in particular when the estimates are based on raw scores (Eckes, 2015; McNamara et al., 2019). Therefore, looking merely at the rater severity main effect without regard to its interaction one may not capture a complete picture of the rater error phenomenon, leading to superficial validity arguments for rater-mediated assessment (Engelhard & Wind, 2018). One effective approach to investigating rater effects is a many-facets Rasch measurement (MFRM) approach (Linacre, 1989) which is capable of detecting both rater main and interaction effects at the group and individual levels more accurately than raw score-based methods (Eckes, 2015; Linacre, 2022).

This study aimed to explore the quality control of real-world high-stakes classroom assessment practices by examining whether teachers displayed a significant DRS effect towards rating criteria and student subgroups (classroom, proficiency, and gender) in an EFL classroom writing examination. This

examination was considered high stakes because the scores primarily informed grading decisions, which had significant consequences for students. The present study extended the existing body of research on rater behaviours and effects by providing more insights into Thai EFL teachers' rating effect, validity, and fairness in a classroom writing assessment. This study also proposed implications for researchers and practitioners with respect to the nature of teachers' rating behaviors, the maintenance of quality control, and the investigation of psychometric quality in the context of a rater-mediated classroom language assessment.

2. Literature Review

2.1 Differential Rater Severity

Performance assessment typically involves constructed-response tasks which require examinees to produce written or spoken performances scored by one or more raters using distinct types of rating scales or rubrics (Knoch, Deygers, & Khamboonruang, 2021; Knoch, Fairbairn, & Jin, 2021). In this way, the scoring of examinee performance is highly subjective by nature, depending inextricably on how well raters are able to interpret and apply scoring rubrics as intended (Knoch, Fairbairn, & Jin, 2021). Unfortunately, raters tend to exercise various forms of judgemental effects or errors which threaten the validity and fairness of ratings (Knoch & Chapelle, 2018). Amongst the varying rater effects, severity is deemed as the most pervasive, persistent, and serious error that needs to be investigated and minimized to ensure rating validity and fairness (Eckes, 2015; Myford & Wolfe, 2003). A rater is considered as manifesting a severity effect when consistently assigning higher or lower ratings on average than those given by other raters (Myford & Wolfe, 2003). Ideally, a rater's severity should be invariant over sources irrelevant to examinee performance or ability (e.g., gender and age) to maintain quality ratings (Engelhard & Wind, 2018). However, when a rater's severity is not systematically invariant towards certain subgroups of test-takers with certain construct-irrelevant characteristics, the rater exhibits a DRS effect and estimates of examinee ability are not comparable between subgroups, posing a threat to the

validity and fairness of ratings (Wind & Guo, 2019). To ensure valid and fair ratings, estimates of rater severity and ratee ability must be invariant over different levels of any undesirable construct-irrelevant factors (Engelhard & Wind, 2018).

2.2 Previous Research

Although a well-developed rubric and substantial rater training can help mitigate rater error, a body of research has well established that raters were prone to exert differing levels of severity (e.g., Khamboonruang, 2020; Youn, 2018) and varying patterns of DRS towards assessment-related facets, for instance, rating criteria (Eckes, 2005; Kondo-Brown, 2002; Schaefer, 2008; Wind & Engelhard, 2013), writing genre (He et al., 2013), task type (Eckes, 2005), task difficulty (Weigle, 1999), and time of rating (Lamprianou et al., 2021). Raters were also inclined to exhibit DRS towards particular subgroups of examinees, for example, age (Wind & Sebok-Syer, 2019), gender (Eckes, 2005; Engelhard & Myford, 2003; Erman Aslanoğlu & Şata, 2021; Wind & Sebok-Syer, 2019), ethnicity (Engelhard & Myford, 2003), proficiency (Engelhard & Myford, 2003; Erman Aslanoğlu & Şata, 2021; Kondo-Brown, 2002; Youn, 2018), and best language (Eckes, 2005). Moreover, rater severity and DRS tend to vary according to such rater characteristics as rater experience (Barkaoui, 2011; Johnson & Lim, 2009; Mohd Noh & Mohd Matore, 2022; Weigle, 1998, 1999; Winke et al., 2013) and rater training (Kang et al., 2019; Mohd Noh & Mohd Matore, 2022; Weigle, 1998, 1999).

Specifically in L2 writing assessments, previous studies have reported mixed findings about DRS effects towards rating criteria and examinee subgroups. Regarding rater-criteria interaction, Kondo-Brown (2002) found that certain native-Japanese raters showed DRS towards content, vocabulary, and mechanics, but showed no DRS towards organization and language use. Schaefer (2008) also found that some raters exhibited different patterns of DRS over scoring criteria associated with content, organization, language use, and mechanics. Similarly, Wind and Engelhard (2013) discovered that raters' severity levels were not invariant over convention, idea, organization, and style. Concerning rater-gender

interaction, Engelhard and Myford (2003) and Eckes (2005) discovered that whilst raters showed no group-level DRS related to student gender, certain raters tended to consistently assign higher or lower ratings than expected to gender subgroups. Erman Aslanoğlu and Şata (2021) revealed, however, that raters did not display both group- and individual-level DRS when scoring Turkish academic writing. In terms of rater-proficiency interaction, Kondo-Brown (2002) found that native-Japanese raters' severity levels were not invariant across L2 native-English students especially whose ability was extremely high or low in Japanese L2 writing assessments. Likewise, Schaefer (2008) reported that some native-English raters were systematically more severe than expected when rating higher-ability Japanese undergraduates but were systematically more lenient than expected when marking lower-ability students in L2 English writing assessment. Other studies reported that raters exercised DRS towards task type (Eckes, 2005; Han, 2021), essay topic and genre (He et al., 2013; Weigle, 1999), examinee ethnicity (Engelhard & Myford, 2003), and examinee best language (Eckes, 2005).

It can be argued from the existing findings that although seemingly self-consistent and invariant in the levels of severity, raters were inclined to demonstrate mixed and varied patterns of DRS towards certain assessment facets and examinee subgroups. Accordingly, scrutinizing solely the severity main effect without its interaction or differential one may fail to capture a thorough rater error, resulting in superficial or even spurious validity arguments. In addition, since performance scores are inextricably rater-mediated, it is thus crucial to systematically investigate the quality of ratings to ascertain that performance ratings are meaningfully interpreted and used in line with intended assessment purposes (Kane, 2013; Knoch & Chapelle, 2018). Notwithstanding numerous findings pertaining to rater main and interaction effects, little is known about such effects, particularly rater-classroom interaction, in the classroom assessment context, and in particular research on Thai EFL raters' behaviors and effects is sparse in the literature. All this necessitates further research into Thai EFL raters' rating behaviors and effects in the classroom assessment context.

E-ISSN: 2287-0024

2.3 Current Research

To shed novel light on Thai EFL classroom teachers' severity and DRS effects, the current research was set out with two specific aims in mind: (1) to investigate teachers' severity variability and (2) to investigate teachers' severity invariance in the context of a high-stakes Thai EFL classroom writing examination. Building upon a MFRM approach, this study conducted a comprehensive DRS analysis with an emphasis on a two-way interaction effect between teacher raters and rubric criteria, student proficiency levels and student genders which were typically found to cause DRS in previous research. Also of particular interest was to ascertain whether individual teachers may have exercised DRS across student classrooms, which has probably remained under-researched. To this end, this study aimed to address two research questions: *(1) Do teachers maintain a uniform level of severity when scoring students' essays in a high-stakes EFL classroom writing assessment?* and *(2) Do teachers exercise DRS over rating criteria, student classrooms, student ability levels, and student genders?* The first research question examined rater severity variability via a three-facet MFRM analysis, whereas the focal second question built on the first research question to investigate whether any of the teachers showed psychometric evidence of DRS towards the rubric criteria and student subgroups in question through a two-way interaction MFRM analysis.

3. Methodology

3.1 Participants and Context

Participants were three classroom teachers and 42 English-major undergraduates (15 males and 27 females) from three English composition classrooms in a public university setting in Thailand. The classrooms were conducted almost entirely online due to the COVID-19 outbreak and were taught by three teachers holding a PhD related to the English language. The first classroom had 14 students (1-14) taught by a male teacher (T1) with about 20 years of teaching experience. The second classroom consisted of 15 students (15-

29) taught by a female teacher (T2) with about 25 years of teaching experience. The final classroom was composed of 13 students (30-42) taught by a male teacher (T3) with about seven years of teaching experience. The students took a three-hour onsite examination deemed as high-stakes since the exam scores were mainly used to inform the teachers' consequential decision about the students' grading. During the examination, the students were asked to write a five-paragraph opinion essay of about 500 words on the same single topic in Microsoft Word using their own notebook.

3.2 Rubric and Rating Procedures

The exam essays were scored using a new analytic rubric developed by the teachers. During a two-hour rubric development session, the teachers together read the course syllabus and existing rubrics (including previous teacher-made rubric, TOEFL iBT holistic rubric, and IELTS analytic rubric), whilst at the same time discussing which criteria should be included in the new rubric and drawing the wording of the criteria in order to construct the first-draft rubric. Following this, the teachers preliminarily trialled the draft rubric to pilot-rate examples of student essays and then revised the draft rubric. After the rubric revision, the teachers discussed and negotiated disagreements before finalizing the rubric. It could thus be said that the rubric was informed mainly by existing scale, curriculum, and intuition (Knoch, Deygers, & Khamboonruang, 2021). The finalized rubric (see Appendix) comprised five criteria or writing ability domains: (1) Thesis and Topic Sentence, (2) Idea Unity and Connection, (3) Idea Development, (4) Vocabulary, and (5) Overall Language. The criteria were rated on a four-point rating scale (2, 3, 4, and 5).

Table 1 summarizes the characteristics of the ratings assigned by each teacher across the students, classrooms, and criteria. The teachers agreed to use a cross-classroom rating design to ensure fair ratings and rated the essays at their convenience. For the student grading, T1 and T3 independently rated Students 1-14 (T1 students), T1 and T2 independently rated Students 15-29 (T2 students),

and T2 and T3 independently rated Students 30-42 (T3 students). For the current research purpose, the teachers were asked to independently rate more randomly selected essays. That is, T1 rated Students 32, 36, 37, 39, 40, 41, and 42 (T3 classroom students), T2 scored Students 1, 4, 5, 7, 11, and 12 (T1 classroom students), and T3 marked Students 15, 17, 18, 20, 23, 26, and 28 (T2 classroom students). Accordingly, the current cross-classroom rating was deemed as an incomplete or partially-crossed rating design in which each student essay was not rated by all teachers (Eckes, 2015; Wind & Guo, 2019), which is practical and common in the classroom context. Despite the partially-crossed rating design, it did still make sense under a MFRM framework to investigate a rater-examinee interaction effect (J. M. Linacre, 2022, personal communication, October 16, 2021) and a MFRM analysis generated robust results based on missing or incomplete rating data (Eckes, 2015; Linacre, 2022).

Table 1

Characteristics of the Ratings Assigned by Three Teachers

Raters	Number of rated criteria	Number of rated students			Total ratings
		Class 1	Class 2	Class 3	
T1	5	14	15	7	180
T2	5	6	15	13	170
T3	5	14	7	13	170

3.3 Data Analysis

To investigate rater severity and DRS effects, this study employed a many-facets Rasch measurement (MFRM) approach, advanced by Linacre (1989) as an extension to the family of Rasch psychometric models (Rasch, 1960). The MFRM offers several advantages over traditional psychometric methods for validating rater-mediated assessment and examining rater behaviors and effects in general and differential rater functioning in particular (Eckes, 2019). Drawing upon the MFRM estimates, it is possible to investigate the main and interaction effects of assessment facets (e.g., rater severity, examinee ability, and rubric difficulty) and

particularly to detect whether a rater differentially rates certain scoring criteria or examinee subgroups more severely or less severely than he or she usually does (Eckes, 2019).

The data were analyzed via the FACETS program (Version No. 3.84.0; Linacre, 2022). In the current MFRM analysis, the Andrich rating scale model was used since the structure of the four-point rating scale was assumed to be the same for all rating criteria. The maximum likelihood method was employed to calibrate the ratings, assigned by three teachers to 42 students' essays across five criteria, for the purpose of estimating the parameters of the latent variables (rater severity, student ability, and criterion difficulty) on the logit scale. The student facet was allowed to vary along the logit scale and was positively oriented, whereas the rater and rubric facets were centred at 0 and were negatively oriented on the logit scale. The MFRM analysis was conducted over two main stages. To begin with, a three-facet MFRM was conducted to investigate rater behaviors, rating scale functioning, and student writing ability. Building on the first stage, a two-way DRS analysis (called bias or interaction analysis in FACETS) was subsequently conducted to investigate significant group- and individual-level DRS between (1) rater severity and rating criteria and (2) rater severity and student subgroups (classroom, gender, and ability) which were created as dummy facets anchored at 0 on the logit scale.

4. Results

The results were organized into three parts. First, the data-model fit Rasch assumption was inspected to ensure meaningful interpretation of the MFRM results. Second, the estimates of the rater severity, student ability, and criterion difficulty were examined with a particular emphasis on investigating whether the teachers were uniform in their levels of severity. Finally, the two-way interaction DRS results were scrutinized to examine whether each of the teachers maintained severity invariance or might have rated certain criteria and student subgroups more or less severely than expected.

4.1 Data-Model Fit

The data-mode fit was inspected based on the Pearson chi-square goodness of fit statistic, the percentage of the unexpected standardized residuals outside ±2 and ±3, and the information-weighted mean square residuals (Infit MnSq) fit statistics of the raters and criteria. The results showed that the chi-square statistic was not significant ($p = .97$), and the Infit MnSq indices of all raters (M = 1.00, SD = 0.30) and criteria (M = 1.00, SD = 0.17) fell within the recommended bounds of 0.50 and 1.50 (Linacre, 2022). Of 520 valid responses, 28 (5.38%) represented the unexpected standardized residuals outside ±2, very close to the expected maximum of 5% (Linacre, 2022), and only 1 (0.19%) accounted for those outside ±3, far below the expected maximum of 1% (Linacre, 2022). All the statistics showed desirable indices, thereby confirming a satisfactory fit of the current data to the Rasch model.

4.2 Rater Severity Variability

Figure 1 displays a variable map showing the levels of the teacher severity, student ability, and criterion difficulty on the common equal-interval logit scale in the first column. Higher-than-zero positive logits indicate higher levels of severity, ability, and difficulty, whilst lower-than-zero negative logits represent lower levels of severity (or higher leniency), ability, and difficulty. The map should be interpreted in conjunction with the group-level statistics in Table 2 and individual-level statistics in Table 3 which yield information about the rater behavior, student ability, and rubric functioning at the group and individual levels, respectively.

Overall, the map displays a wide spread of rater severity, student ability, and criterion difficulty logits. Interestingly, the distribution of the student ability logits was lower than that of the rater severity and criterion difficulty logits, implying that most of the students received low scores assigned by the teachers across the criteria. The rating score categories (2, 3, 4, and 5) in the fifth column were in a desired hierarchical order, where higher scores, which were more difficult and require higher ability to achieve, were placed higher than lower scores (Linacre, 2022). The length of Score 5 was very narrowed on the logit scale, meaning that it

was rarely assigned by the teachers to the student essays. In other words, only a small number of the student essays were judged as satisfying the 5-point quality description across five ability domains. On the other hand, Score 3 showed the largest proportion, implying that it was most frequently used and most of the students' essays were evaluated as fulfilling the three-point quality description across five ability domains.

In line with the map, the separation and fixed chi-square statistics in Table 2 confirm a significant difference in the teachers' severity logits. The significant fixed chi-square statistic ($p < .05$) indicated that at least two of the teachers differed significantly in severity (Linacre, 2022). The rater separation ratio of 3.86 was greater than the expected value of 1, suggesting that the teachers' severity was not uniform (Eckes, 2015). The rater separation strata of 5.48 indicated that the teachers' severity levels could be stratified into about five statistically distinct classes (Eckes, 2015). The high rater reliability (0.94) of the separation statistics confirms significant variations in the teachers' severity levels (Eckes, 2015). As shown in Table 3, T2 showed the highest severity at 0.52 logits, which was also close to that of T1 at 0.13 logits, whereas T3 exhibited the lowest severity at -0.65 logits. The rater Infit MnSq indices were close to the expected index of 1 and within the acceptable range (0.50 and 1.50), indicating that each teacher was self-consistent in his or her level of severity on the whole (Linacre, 2022). The small standard errors of estimate ($SE = 0.15$) close to 0 suggests a precise estimation of the rater severity logits (Linacre, 2022).

Regarding the rubric functioning, the significant fixed chi-square statistic ($p < .05$), together with the criterion separation ratio (3.45), strata (4.93), and reliability (0.92) altogether suggested varying difficulty levels of the criteria that could be stratified very reliably into almost five statistically distinct classes (Linacre, 2022). Amongst the rubric criteria, Idea Development and Overall Language were difficult (logit = 0.80) and Overall Language showing the second-highest difficulty (logit = 0.69). Thesis and Topic Sentence, Vocabulary, and Idea

Unity and Connection were less difficult, with the logits of -0.68, -0.43, and -0.39, respectively. All the criterion Infit MnSq values were also acceptable, indicating that, on the whole, each criterion was consistently assigned ratings by the teachers for the current group of students (Linacre, 2022).

Regarding student ability, the significant fixed chi-square statistic ($p < .05$), along with the student separation ratio (1.81), strata (2.75), and reliability (0.77) all suggested that the ability levels of this group of 42 students could be stratified relatively reliably into almost three statistically distinct classes, implying the teachers and rubric could differentiate the quality of the student essays. Due to limited space, the students were sub-grouped according to their classrooms, ability levels, and genders for further DRS investigation. The students with logits above -1.00, between -1.00 and -2.00, and below -2.00 on the logit scale were grouped as high-, mid-, and low-ability students, respectively.

E-ISSN: 2287-0024

Figure 1

Variable Map

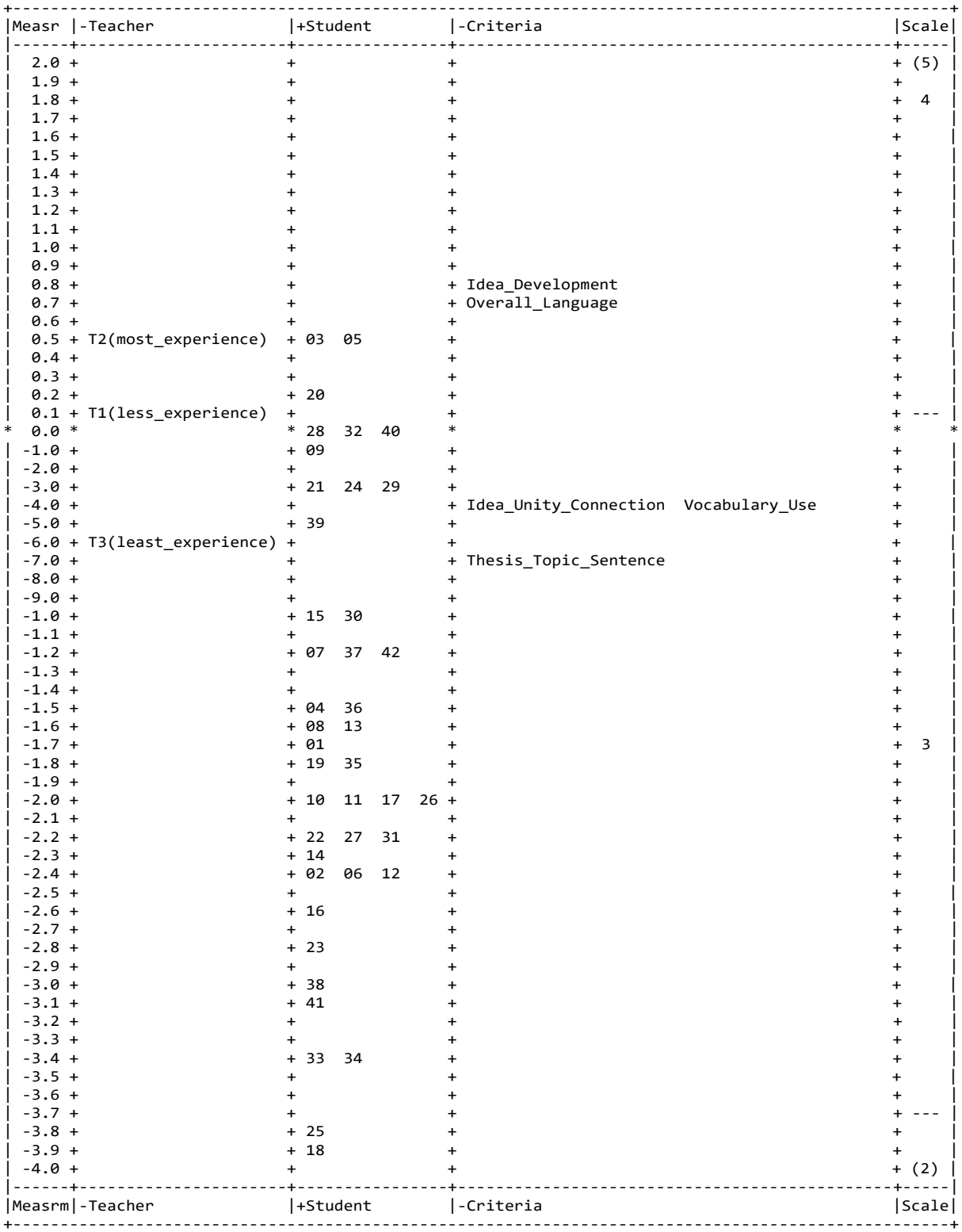


Table 2

Group-Level Statistics

Statistics	Rater	Criteria	Student
Separation ratio	3.86	3.45	1.81
Separation strata	5.48	4.93	2.75
Separation reliability	0.94	0.92	0.77
Fixed Chi-square	$p = .00$	$p = .00$	$p = .00$

Table 3

Individual-Level Statistics

Facet elements	Rating		Average		Estimate		Infit
	Score	Count	Observed	Fair	Logit	*SE*	MnSq
T2 (most severe)	493	170	2.90	2.91	0.52	0.15	1.03
T1 (less severe)	548	180	3.04	3.00	0.13	0.15	0.97
T3 (least severe)	547	170	3.22	3.19	-0.65	0.15	0.99
ID (most difficult)	296	104	2.85	2.84	0.80	0.20	1.09
OL (very difficult)	299	104	2.88	2.87	0.69	0.20	1.26
UC (relatively easy)	328	104	3.15	3.12	-0.39	0.19	0.71
VC (relatively easy)	329	104	3.16	3.13	-0.42	0.19	0.81
TT (easiest)	336	104	3.23	3.20	-0.68	0.19	1.13
T1Class students ($n = 14$)	37.86	12.14	3.12	3.09	-1.44	0.58	0.94
T2Class students ($n = 15$)	36.93	12.33	2.98	3.03	-1.65	0.57	1.01
T3Class students ($n = 13$)	38.77	12.69	3.02	3.01	-1.70	0.56	0.97
High-ability students ($n = 12$)	44.33	12.92	3.43	3.43	-0.09	0.53	1.19
Mid-ability students ($n = 12$)	38.42	12.50	3.07	3.05	-1.51	0.57	0.99
Low-ability students ($n = 18$)	33.06	11.94	2.76	2.77	-2.67	0.59	0.82
Male students ($n = 15$)	38.80	12.67	3.07	3.07	-1.47	0.56	1.07
Female students ($n = 27$)	37.26	12.22	3.02	3.02	-1.67	0.57	0.92

Note. ID = Idea Development; OL = Overall Language; UC = Idea Unity and Connection; VC = Vocabulary; TT = Topic and Thesis Sentence

4.3 Group-Level Differential Rater Severity

Table 4 presents group-level DRS instances. The analysis yielded a total of 15 teacher-criteria interactions, nine teacher-classroom interactions, six teacher-

gender interactions, and nine teacher-proficiency interactions that showed possible DRS. Due to space limitation, only significant DRS cases were presented.

Table 4

Group-Level Significant Interactions

Interaction pair		Rating count	Observed score	Expected score	Obs-Exp average	Bias size	Infit MnSq	Significance test		
								df	*t*	*p*
T1	ID	36	89	102.23	-0.37	-1.53	0.9	35	-4.29	.0001
T1	VC	36	121	113.62	0.21	0.77	0.6	35	2.40	.0217
T1	OL	36	110	103.27	0.19	0.73	0.5	35	2.23	.0328
T3	OL	34	92	103.24	-0.33	-1.30	1.1	33	-3.78	.0006
T3	ID	34	113	102.24	0.32	1.20	0.5	33	3.65	.0009
T3	VC	34	107	113.21	-0.18	-0.69	0.5	33	-2.05	.0485
T1	T1Class	70	228	213.61	0.21	0.79	0.8	69	3.41	.0011
T1	T3Class	35	103	110.09	-0.20	-0.79	0.8	34	-2.33	.0259
T3	T3Class	65	216	207.06	0.14	0.51	0.7	64	2.14	.0358
T3	T1Class	70	217	227.70	-0.15	-0.59	0.8	69	-2.50	.0149
T3	Male	60	185	193.10	-0.14	-0.51	1.1	59	-2.03	.0471

In respect of the teacher-criterion interaction, the significant fixed chi-square statistic indicated that there was statistically significant DRS between 15 teacher-criterion interactions, $X^2(15)$ 68.7, $p = .00$. All the Infit MnSq values in the final column were between 0.50 and 1.50, indicating that individual teachers were consistent in their DRS patterns across the criteria. Of the 15 interactions, only T1 and T3 showed reversed patterns of significant DRS towards Idea Development, Vocabulary, and Overall Language, totalling six significant interaction effects. For example, T1 rated Idea Development 36 times or counts which made up a total observed score of 89. Yet, based on T1 overall severity and the overall difficulty of Idea Development, the expected score of Idea Development should be 102.23, far over the observed score by -13.03. The average score of -0.37 (89 – 102.23 ÷ 36 = -0.37) indicated that the observed score was -0.37 points on average much lower than it might have been expected, making up the overall DRS or bias size of -1.53 logits which was also statistically significant ($t > -2.00$, $p < .01$). A bias size over 0.50 is typically deemed as strong and serious (Isbell, 2017, p. 4). Therefore, T1

was significantly more severe than he usually was by -1.53 logits and the Infit MnSq of 0.9 suggested T1 was consistently more severe than expected across all the students judged towards Idea Development. Overall, T1 was more severe towards Idea Development than he was on average by -1.53 logits and less severe towards Vocabulary and Overall Language than he was overall by 0.77 and 0.73 logits, respectively. Conversely, T3 was less severe than expected towards Idea Development by 1.21 logits and more severe than expected towards Vocabulary and Overall Language by -0.69 and -1.30 logits, respectively. Interestingly, T2 did not appear to show significant DRS and therefore was invariant in her severity level across the criteria.

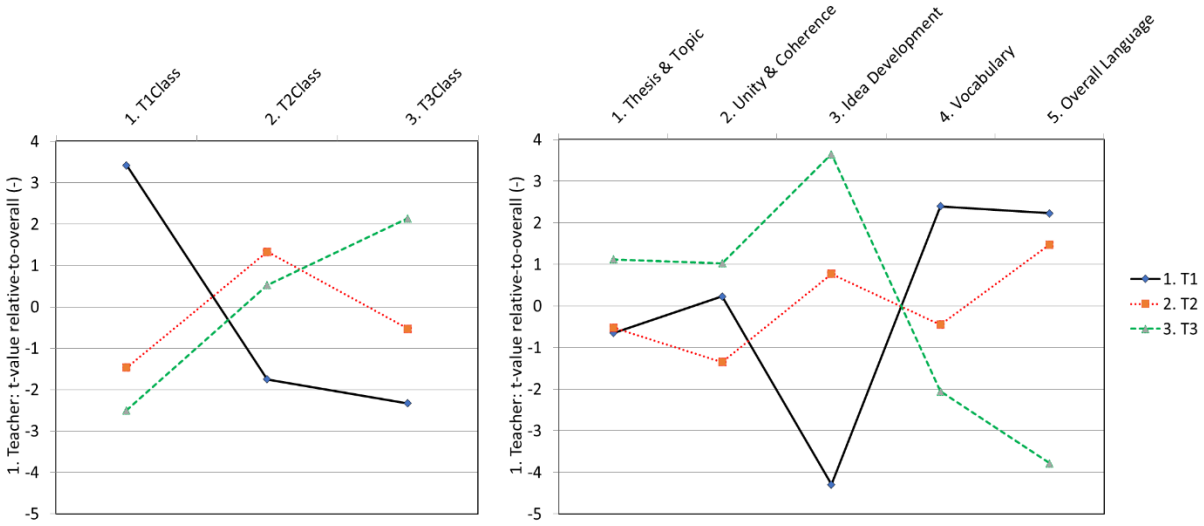
Regarding the teacher-classroom interaction, the significant fixed chi-square test indicated significant DRS on the whole, $X^2(9)$ 35.5, $p = .00$. On closer inspection, four interactions showed statistically significant bias size ($t > \pm 2$, $p < .05$), which was exhibited by T1 and T3 interactions with T1Class and T3Class students. This suggests that T1 rated his classroom students less severely but scored T3 classroom students more severely than expected, whereas T3 rated his classroom students less severely but scored T1 classroom students more severely than expected. Interestingly, the most severe and experienced teacher (T2) did not show significant DRS towards any classrooms. With respect to the teacher-ability interaction, the non-significant fixed chi-square statistic suggested no significant DRS between the teachers and student ability groups, $X^2(9)$ 1.4, $p = 1.00$. As regards teacher-gender interaction, although the non-significant fixed chi-square statistic confirmed no significant overall DRS, $X^2(6)$ 11.2, $p = .08$, T3 was significantly more severe than he was on average towards male students.

Figure 2 displays a line graph showing the teachers' overall DRS patterns across student classrooms and rubric criteria based on *t*-statistic values presented in Table 4. The *t*-statistic values outside positive and negative 2 suggest that the teachers are significantly less severe or more severe, respectively, than expected. As can be seen, T1, for example, was significantly more lenient than expected

towards Vocabulary, Overall Language, and his classroom students, but significantly more severe than expected towards Idea Development and T3 classroom students. By contrast, T3 showed a reverse pattern of DRS.

Figure 2

Line Graph Showing Teacher Group-Level DRS Patterns



4.4 Individual-Level Differential Rater Severity

Table 5 reports only individual-level significant DRS cases across rating criteria and student subgroups. Originally, the analysis generated a total of 51 interactions (30 teacher-criteria interactions, nine teacher-classroom interactions, three teacher-gender interactions, and nine teacher-proficiency interactions) that signalled feasible DRS. Owing to limited space, only 17 statistically significant DRS cases were presented in the table. Significant interactions suggest that individual raters assigned ratings either significantly higher or lower than the overall group of raters (Wind & Engelhard, 2013). Take T1 for example, he judged Idea Development at 1.66 logits but rated Vocabulary at -0.64 logits and therefore was 2.29 logits more severe on Idea Development than with Vocabulary, which was also statistically significant ($t > 2.00$, $p < .05$). Overall, there were 12 significant teacher-criteria interactions, three significant teacher-classroom interactions, and two significant teacher-gender interactions. T1 showed significant DRS over six pairs of criteria and two pairs of classrooms, T3 exerted significant DRS over six

pairs of criteria, one pair of classrooms, and one pair of genders, and T2 exhibited significant DRS towards only one pair of genders. Taken both significant and non-significant interactions together, all teachers surprisingly showed DRS proclivity in favor of their own classroom students even though only T1 and T3 exhibited reversed patterns of significant DRS towards their own and each other's classroom students. That is, T1 was systematically more severe with T3 students than with his students, whereas T3 was more severe with T1 students than with his students. T2 and T3 also demonstrated reverse patterns of DRS towards student gender subgroups. No significant DRS was found towards student ability subgroups, implying that the teachers' severity was invariant in differentiating students' essay quality. Overall, T1 and T3 each displayed about 47% and T2 about 6% of the 17 significant DRS instances.

Table 5

Individual-Level Significant Interactions

Target teacher	Context source	Target logit	Obs-Exp average	Context source	Target logit	Obs-Exp average	Contrast logit	Rasch-Welch		
								df	*t*	*p*
T1	ID	1.66	-0.37	VC	-0.64	0.21	2.29	69	4.80	.0000
T1	ID	1.66	-0.37	OL	-0.60	0.19	2.25	69	4.66	.0000
T1	UC	0.05	0.02	ID	1.66	-0.37	-1.60	69	-3.33	.0014
T1	TT	0.34	-0.06	ID	1.66	-0.37	-1.32	69	-2.73	.0080
T1	TT	0.34	-0.06	VC	-0.64	0.21	0.98	69	2.15	.0352
T1	TT	0.34	-0.06	OL	-0.60	0.19	0.94	69	2.04	.0457
T3	ID	-1.85	0.32	OL	0.65	-0.33	-2.50	65	-5.25	.0000
T3	ID	-1.85	0.32	VC	0.04	-0.18	-1.89	65	-4.02	.0002
T3	TT	-1.01	0.10	OL	0.65	-0.33	-1.66	65	-3.53	.0008
T3	UC	-0.99	0.09	OL	0.65	-0.33	-1.64	65	-3.45	.0010
T3	TT	-1.01	0.10	VC	0.04	-0.18	-1.05	65	-2.26	.0272
T3	UC	-0.99	0.09	VC	0.04	-0.18	-1.02	65	-2.19	.0323
T1	T1Class	-0.66	0.21	T3Class	0.92	-0.20	-1.58	66	-3.86	.0003
T1	T1Class	-0.66	0.21	T2Class	0.53	-0.10	-1.20	142	-3.66	.0004
T3	T1Class	-0.06	-0.15	T3Class	-1.16	0.14	1.10	132	3.28	.0013
T2	Male	0.08	0.11	Female	0.78	-0.06	-0.70	122	-2.19	.0303
T3	Male	-0.13	-0.14	Female	-0.92	0.07	0.79	119	2.51	.0134

Note. ID = Idea Development; OL = Overall Language; UC = Idea Unity and Connection; VC = Vocabulary; TT = Topic and Thesis Sentence

5. Discussion

The present study applied a many-facets Rasch measurement (MFRM) technique to investigate classroom teachers' severity and differential rater severity (DRS) effects in a local high-stakes Thai EFL classroom writing examination. The DRS focused on a two-way interaction effect between rater severity and analytic rubric criteria, and between rater severity and student subgroups (classroom, gender, and ability). This study was probably the first to explore DRS towards student classroom. The current research discovered several interesting findings for the research questions and further discussions.

As regards the first research question, the present findings revealed that the teachers were not uniform in their severity levels, too much to be acceptable for a high-stakes classroom examination. This supported previous research reporting that EFL teachers still exerted high-severity variability upon receiving rater training (e.g., Khamboonruang, 2020, 2022) and that even certified raters still exhibited varied severity in standardized testing contexts (Eckes & Jin, 2021). The findings also indicated that the most experienced teacher was the most severe and the least experienced teacher was the most lenient, supporting some findings that more experienced raters exercised higher severity (Barkaoui, 2011; Mohd Noh & Mohd Matore, 2022), and contradicting certain findings that less experienced raters exercised more severity in writing assessment (Weigle, 1998, 1999). In fact, the teachers' severity variability distorted the raw score-based estimates of the students' writing ability and resulted in inaccurate and unfair grading which was deemed as high-stakes. If the students' ability estimates had been based on MFRM or in other words had the students' ratings been adjusted for the teachers' severity differences, the rater severity variability would not have distorted the ability estimates or some students' grades would have been different from the one they received, precisely because MFRM-based ability estimates are corrected for variations in the rater severity (Eckes, 2015; McNamara et al., 2019).

In relation to the second research question, the current findings revealed that the most experienced and severe teacher demonstrated only a small proportion of significant DRS and was therefore largely invariant in severity in the classroom assessment, whereas the less experienced and less severe teachers equally exhibited a substantial amount of significant DRS and were thereby slightly invariant in severity. This partly supported and contradicted Erman Aslanoğlu and Şata's (2021) study revealing that not only severe raters but also lenient raters were prone to exercise significant DRS. Interestingly, the less experienced teachers displayed reverse patterns of significant DRS across the analytic criteria and student classrooms, supporting previous findings in terms of reverse and idiosyncratic patterns of DRS (Kondo-Brown, 2002; Youn, 2018). Of all the interactions, the teachers displayed a great deal of DRS towards rubric criteria and student classrooms, and an observed amount of DRS towards student genders. Evidence of large DRS over criteria was also typical in previous research (Eckes, 2005; Han, 2021; He et al., 2013; Kondo-Brown, 2002; Schaefer, 2008; Wind & Engelhard, 2013; Youn, 2018). The predominant DRS toward the criteria implied that the teachers' severity may not be invariant over certain criteria and/or they may not be congruent in their interpretation of certain criteria. However, teacher-criteria DRS may not be of grave concern as was pointed out by Eckes (2005) that DRS across scoring criteria is less of a problem, for rating criteria are aimed at measuring different domains of language performance. Of all the criteria, only Idea Development, Vocabulary, and Overall Language were systematically rated more or less severely than expected, in line with previous research showing that raters tended to show DRS toward grammar-related criteria (He et al., 2013; Schaefer, 2008) which were considered as similar to Overall Language in the current rubric where linguistic errors and accuracy were taken into account. Yet, Kondo-Brown (2002) revealed no significant DRS on Language Use. The most interesting finding from this research is probably that the less experienced teachers rated their own classroom students less severely but scored each other's classroom students more severely than they typically did, which has probably never before been unveiled in any prior research. In fact, rater-classroom DRS exerts a more serious threat to

the validity and fairness of classroom assessment than teacher-criteria DRS which is, by nature, difficult to avoid, even in well-designed and standardized assessment contexts. Interestingly, some teachers showed significant DRS towards student genders, which both supported previous research (e.g., Eckes, 2005; Weigle, 1999) and contradicted prior research (e.g., Erman Aslanoğlu & Şata, 2021). Additionally, whilst previous research found significant DRS towards students of different proficiency levels (Erman Aslanoğlu & Şata, 2021; Schaefer, 2008; Youn, 2018), no significant DRS was detected towards student ability in this study, implying that despite showing DRS towards different classrooms, the teachers were homogeneous in differentiating the students' essay quality.

There are several plausible explanations for the teachers' variations in the levels of severity and the patterns of DRS in the classroom context. The major underlying factor would be their different levels of teaching experience. Whilst the current findings detected a systematic relationship between teaching experience and severity and DRS effects, it should be realized that more teaching experience may not necessarily guarantee more rating experience and quality rating and it remains unclear as to whether the teachers' teaching experience or age or perhaps both underlay variations in the severity and DRS effects. Other rater-related and contextual variables might also have contributed directly towards the teachers' severity and DRS effects, and/or indirectly influenced the relationship between teaching experience and severity and DRS effects. This calls for further research to comprehensively investigate factors affecting rater effects in order to capture a complete relationship between severity and DRS effects and their contributory factors in the classroom context. Apart from teaching experience, it might be plausible that the rubric criteria developed by the teachers were not clear enough and thus needed further revision. The varying linguistic features of the student essays with different quality levels might also have affected the teachers' decision-making process. Another factor that might have affected the teachers' scoring decision was grading and assessment policy since in the classroom context, test scores assigned to students were not just based purely on their

language proficiency per se but may also be influenced by grading/assessment policy implemented in the context. Indeed, the grading policy may have directed the way in which the teachers assigned rating scores and tailored grades in the classroom assessment. Teacher-student relationship may also have influenced the teachers' decision to assign scores as they might set up their expected minimum score to prevent the students from failing the course. Other plausible factors would be insufficient training and other contextual factors (e.g., teacher workload and rating conditions) within the ongoing classroom which tend to be highly varied by nature. What is intriguingly more difficult to explain is why some teachers exhibited DRS in favour of their own classroom students but against other teachers' classroom students. One of the feasible underlying reasons of this might be that since the scores also reflect the quality of teaching performance, they might not have wanted their students to get low scores. Additionally, when the teachers saw that his or her classroom students received low ratings by other teachers, this might have played a part in their high ratings for his or her classroom students. It may thus be argued that a cross-classroom rating may be one example of the DRS causes arising in a classroom assessment situation where teachers cross-rate each other's classroom students for the purpose of ensuring fair rating, which is obviously not the case as evidenced by the present findings. This kind of rating design may introduce unpredictable construct-irrelevant sources of DRS which are detrimental to the quality of a rater-mediated classroom assessment.

The current study has not been without its limitations. Firstly, the number of students in the student subgroups were small, which may have influenced the MFRM estimates and hence the DRS results. In particular, the number of male and female students was different, which may have impacted the teacher-gender DRS results. Another caveat was the inability to control and monitor teachers' rating conditions in the classroom assessment since it was typical for teachers to rate student essays at their convenient time. Finally, although MFRM estimates based on partially crossed ratings were still meaningful and robust, this type of rating data resulted in more errors in the estimates than fully crossed ratings (Eckes,

2015). If a fully crossed rating design had been employed, the current findings might have varied, thus limiting the generalizability of the current findings regarding the DRS effect for students from different classes.

6. Conclusion

Despite certain limitations, the findings from this research offered several insights into classroom teachers' severity and DRS effects (particularly teacher-classroom interaction) and quality control in a rater-mediated high-stakes EFL classroom assessment context. In summary, the current findings revealed that teachers were not homogeneous in the level of severity they exercised even after developing, revising, and trialling the rubric together. Although individual teachers seemed to maintain their severity level over the assessment conditions on the whole, certain teachers were not severely or leniently invariant across rating criteria and student subgroups. Interestingly, less experienced teachers showed reversed DRS patterns and larger DRS than a more experienced teacher. Surprisingly, teachers tended to rate their own classroom students more leniently but score other classroom students more severely. The current findings underscored the idiosyncrasy and convolution of classroom teachers' rating behaviors in terms of severity and DRS effects. Severity and leniency forms of DRS are both a threat to assessment validity and fairness. Without a MFRM-driven DRS analysis, the present study could not have obtained reliable and fine-grained information about classroom teachers' severity and DRS effects.

A number of plausible conclusions could be drawn from the present findings. Firstly, a cross-classroom rating design may or may not really maintain rating validity and fairness in the classroom assessment context. This accentuates the need for new rating designs and more standardized classroom assessment to ensure valid and fair high-stakes assessments, and for consequential decisions to be made based on assessment outcomes. Secondly, rater experience variability influences variations not only in severity but also in DRS. Thirdly, in the classroom assessment context, teachers' rating decisions tend to be influenced by varying

sources of DRS, which may be more complicated and difficult to predetermine and control than those manifesting in other assessment contexts. Finally, a MFRM approach is particularly useful for ensuring the quality control of rater-mediated classroom assessments. A MFRM approach helps identify specific DRS patterns of individual teachers, in turn helping tailor rater training activities to meet their specific needs and helping them to become aware of their rating variance with a view to minimizing rater effects on valid inferences from a student's performance ability and grade. The current findings raise the attention and awareness of teachers, educators, and policymakers concerning the impact of rater effects on the validity and fairness of rater-mediated assessment in the classroom context.

7. Implications

The current findings have implications for rater-mediated classroom writing assessment practice and research. The present findings point out that a cross-classroom rating design introduces serious rating bias, and that teachers are prone to bring unpredictable construct-irrelevant sources into their rating of students' language performance within the context of classroom assessment. To ameliorate teachers' rating bias, teachers should not rate their own classroom students and know each other's rating scores and student identity (e.g., name and classroom) which should be blinded to reduce teachers' bias in rating. However, blinding may not be practical in the context of classroom assessment since teachers typically want to know individual students' score and identity to detect individuals' learning progression and achievement. With recent advances in technology, teachers are highly encouraged to take advantage of emerging classroom management platforms, such as Canvas which has a blind review function for essay ratings, and generative artificial intelligence (AI) tools, such as ChatGPT and Grammarly, to facilitate and streamline classroom writing assessment. If possible, teachers are also highly recommended to utilize automated scoring programs to partly support students' performance scoring, which would help cushion the effect of teachers' scoring bias. Additionally, teachers are advised to evaluate students' performance based on MFRM-generated estimates to ensure more valid and fairer rating scores

and grades. Future scale development or revision and rater training should pay particular attention to the rating criteria towards which raters tend to exhibit varied severity to ensure more valid and fairer ratings. The wording of such criteria needs to be described more clearly and raters need to discuss and practice more on such criteria. More attention, norming, and practice should be focused on the facets towards which teachers are prone to display DRS, especially the rating criteria that are judged unusually more or less severely by teachers.

The current findings point towards a line of research into potential sources of DRS and other forms of differential rater functioning that are worth investigating. Future research should employ a mixed-methods research methodology to gain a more thorough understanding of the rater effect phenomenon. For example, researchers may apply both Generalizability Theory and MFRM quantitative approaches to investigate rater effects, in particular interaction variances between raters and other facets, which could offer a more insightful and comprehensive account of the differential rater functioning and DRS phenomena. Apart from psychometric methods, researchers should employ qualitative techniques (e.g., interview, think-aloud, eye-tracking, and focus group) to delve deeply into teachers' cognitive and meta-cognitive rating strategies and other information from teachers that may divulge potential causes of severity and DRS effects in the classroom assessment context. In tandem with DRS, researchers are encouraged to examine other forms of differential rater functioning, such as differential centrality, and explore whether there exists any systematically interdependent relationship amongst them. Very recently, Jin and Eckes (2022) proposed a dual differential rater functioning model which, they claim, can detect and measure not only differential severity but also differential centrality in which raters' tendency to overuse the rating scale's middle score category or categories is not invariable. Findings from an investigation of multiple forms of differential rater functioning may offer more in-depth and novel insights into the rater effect/error phenomenon, and detailed feedback for improving rater training and monitoring and for raising teachers' awareness of such effects. It

would also be useful to extend the present study by examining the impact of teachers' severity and centrality differences as well as differential severity and centrality effects on student writing scores and grades for various student subgroups. Apart from a two-way interaction effect, a three-way interaction analysis, for example between rater characteristics, assessment tasks, rating scales, and students, could render more insights into the patterns and root causes of DRS. Weigle's (1998, 1999) studies highlighted multiple interactions effects between rater severity, rater experience, rater training, and task difficulty. In Weigle's (1999) study, she found that the level of task difficulty influenced rater variability between inexperienced and experienced raters who showed no significant severity differences on an easy task but significant severity differences on a difficult one, with inexperienced raters appearing more severe on a difficult task. It was beyond the scope of this study to probe into sources underlying the teacher DRS, calling for further research to investigate potential sources that may mediate or moderate the relationship between rater backgrounds and rater effects. This could offer a fuller understanding of the rater effect phenomenon. For validation research, DRS and/or other forms of differential rater functioning can be investigated to provide backing for various aspects of validity arguments, such as decision, explanation, and consequence inferences (Kane, 2013; Knoch & Chapelle, 2018). To reiterate, we need more in-depth studies into rater effects within the classroom assessment context to work out potential sources of and in turn practical solutions to rater effects, in particular a perennial differential rater severity.

8. About the Author

Apichat Khamboonruang (ORCID iD: 0000-0002-7182-3501) is currently a lecturer of English at Chulalongkorn University Language Institute (CULI), Bangkok, Thailand. He holds a Ph.D. in Applied Linguistics from the University of Melbourne, Australia. His research and academic interests are related to language testing and assessment and research methods in applied linguistics.

9. Acknowledgement

I would like to thank Mike Linacre for kindly answering my questions and giving me extremely helpful suggestions about many-facets Rasch measurement and FACETS.

10. References

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

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*(3), 197–221. https://doi.org/10.1207/s15434311laq0203_2

Eckes, T., & Jin, K.-Y. (2021). Examining severity and centrality effects in TestDaF writing and speaking assessments: An extended Bayesian many-facet Rasch analysis. *International Journal of Testing, 21*(3–4), 131–153. https://doi.org/10.1080/15305058.2021.1963260

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed., Vol. 22). Peter Language.

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, G. J., & Myford, C. M. (2003). Monitoring faculty consultant performance in the advanced placement English literature and composition program with a many-faceted Rasch model. *ETS Research Report Series*, i-60. https://doi.org/10.1002/j.2333-8504.2003.tb01893.x

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

Erman Aslanoğlu, A. & Şata, M. (2021). Examining the differential rater functioning in the process of assessing writing skills of middle school 7th grade students. *Participatory Educational Research, 8*(4), 239–252. https://doi.org/10.17275/per.21.88.8.4

Han, C. (2021). Detecting and measuring rater effects in interpreting assessment: A methodological comparison of classical test theory, generalizability theory, and many-facet Rasch measurement. In Chen, J., Han, C. (eds) *Testing and assessment of interpreting*. New Frontiers in Translation Studies. Springer. https://doi.org/10.1007/978-981-15-8554-8_5

He, T.-H., Gou, W. J., Chien, Y.-C., Chen, I.-S. J., & Chang, S.-M. (2013). Multi-faceted Rasch measurement and bias patterns in EFL writing performance assessment. *Psychological Reports, 112*(2), 469–485. https://doi.org/10.2466/03.11.PR0.112.2.469-485

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

Jin, K.-Y., & Eckes, T. (2022). Detecting differential rater functioning in severity and centrality: The dual DRF facets model. *Educational and Psychological Measurement, 82*(4), 757–781. https://doi.org/10.1177/00131644211043207

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing, 26*(4), 485–505. https://doi.org/10.1177/0265532209340186

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

Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing, 36*(4), 481–504. https://doi.org/10.1177/0265532219849522

Khamboonruang, A. (2022). Building an initial validity argument for binary and analytic rating scales for an EFL classroom writing assessment: Evidence

from many-facets Rasch measurement. *rEFLections, 29*(3), 675–699. https://so05.tci-thaijo.org/index.php/reflections/article/view/262690

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

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.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3–31. https://doi.org/10.1191/0265532202lt218oa

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

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

Linacre, J. M. (2022). *Facets computer program for many-facet Rasch measurement* (version 3.84.0.) [Computer Software]. Available from http://www.winsteps.com/

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

Mohd Noh, M. F., & Mohd Matore, M. E. E. (2022). Rater severity differences in English language as a second language speaking assessment based on

rating experience, training experience, and teaching experience through many-faceted Rasch measurement analysis [Original Research]. *Frontiers in Psychology, 13,* 1-13. https://doi.org/10.3389/fpsyg.2022.941084

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.

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

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465–493. https://doi.org/10.1177/0265532208094273

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287. https://doi.org/10.1177/026553229801500205

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145–178. https://doi.org/10.1016/S1075-2935(00)00010-6

Wind, S. A., & Engelhard, G. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing, 18*(4), 278–299. https://doi.org/https://doi.org/10.1016/j.asw.2013.09.002

Wind, S. A., & Guo, W. (2019). Exploring the combined effects of rater misfit and differential rater functioning in performance assessments. *Educational and Psychological Measurement, 79*(5), 962–987. https://doi.org/10.1177/0013164419834613

Wind, S. A., & Sebok-Syer, S. S. (2019). Examining differential rater functioning using a between-subgroup outfit approach. *Journal of Educational Measurement, 56*, 217-250. https://doi.org/10.1111/jedm.12198

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*(2), 231–252. https://doi.org/10.1177/0265532212456968

Youn, S. J. (2018). Rater variability across examinees and rating criteria in paired speaking assessment. *Papers in Language Testing and Assessment, 7*(1), 32–60.

11. Appendix

The Analytic Rubric

Score Category	Rating Criteria or Writing Ability Domains				
	Thesis and Topic Sentence	**Idea Unity and Connection**	**Idea Development**	**Vocabulary**	**Overall Language**
5	The thesis or topic sentence is well-structured, clear, and convincing.	Supporting ideas for each main idea are logically arranged using appropriate transitions, are related to the single main idea, and are not redundant or overlapping.	Supporting ideas are sufficiently provided with appropriate explanation or elaboration and the writer anticipates the reader's arguments and provides at least one counterargument.	The writer uses a variety of words and/or academic and specific words related to the topic and all words or expressions are used appropriately in the context.	The essay shows clear and comprehensible language use and syntactic variety though it may have minor linguistic errors that do not interfere with meaning.
4	The thesis or topic sentence is clear, and convincing. but not well-structured.	Supporting ideas are logically arranged to some extent but few transitions are not appropriately used to connect ideas. Few supporting ideas are not related to the single main idea and are redundant.	Supporting ideas are sufficiently provided with somewhat appropriate explanation or elaboration, but there may or may not have the writer's counterargument.	The writer uses a variety of words and/or academic and specific words related to the topic, but few words or expressions are not appropriately used in the context.	The essay generally shows clear and comprehensible language use and syntactic variety though it occasionally has linguistic errors that do not interfere with meaning.
3	The thesis or topic sentence exists but does not show a clear position of the writer or does not well respond to the prompt.	Many supporting ideas are not logically arranged and many transition signals are not appropriately used to connect ideas. Many supporting ideas are not related to the single main idea and are redundant.	Supporting ideas are not sufficiently provided and the supporting appropriate explanation or elaboration is not convincing.	The writer uses a noticeable variety of words and/or academic and specific words related to the topic, but many words or expressions are not appropriately used in the context.	The essay shows many unclear expressions and/or sentences, lacks syntactic variety, and has many linguistics errors occasionally obscure meaning.
2	The thesis or topic sentence does not exist in the introduction paragraph.	Most supporting ideas are not logically arranged. Transition signals are not used, or most are not used appropriately to connect ideas. Most supporting ideas are not related to the single main and are redundant.	Very few supporting ideas are provided and the supporting explanations, exemplifications and/or details are weak or not convincing.	The writer uses a very limited range of words and/or academic and specific words related to the topic, but most words or expressions are not appropriately used in the context.	The essay shows many unclear expressions and/or sentences, lacks syntactic variety, and, often has many serious linguistic errors that obscure meaning.