Language Testing in Focus: An International Journal

2021, Vol. 4, 23-36

Rater-task interaction effects on testing EFL learners' written performances

Rania Zribi*
Faculty of Letters and Humanities of Sfax, Tunisia
Chokri Smaoui
Faculty of Letters and Humanities of Sfax, Tunisia

Received: 02 April 2021

Accepted: 19 November 2021

Abstract

This study examines the interaction effect between rater groups and tasks on evaluating EFL learners' written performances. Fifty raters took part in this study. The experienced rater group (n=25) and the novice rater group (n=25) judged sixty essays (30 narratives and 30 argumentative writing modes) written by third-year English students. Raters' decision-making behaviours, in terms of scores assignment and written comments, were diagnosed based on different quantitative and qualitative tools. Scores were analysed based on FACETS to examine the effects of rater-task interaction on raters' severity and internal consistency and the analytic scale's functionality. Qualitative data were also analysed to diagnose which aspects of writing were deemed more important than others across rater groups and task types. The analysis revealed that both raters and tasks were substantially influential factors. The majority of expert raters displayed more severity in assessing narrative essays than argumentative essays. Different qualitative judgments are also detected across raters and tasks due to rating experience and task requirements. The findings of this study reflected implications not only for testing learners' writing proficiency but also for test validation research in the task-based writing performance assessment field.

Keywords: Raters, tasks, interaction effects, scores' variation, severity, feedback, scores

Introduction

As a formal communicative skill, writing plays a pivotal role in the community where people communicate by exchanging opinions and conveying their messages and sharing information. This goes in line with Caswell and Mahler's (2004, p. 3) view that writing is a crucial interactive mode useful in all aspects of social life. For instance, doctors write prescriptions to their patients and friends type messages or emails to communicate with each other. Because people communicate with each other for a wide variety of purposes, their performances are produced in many different forms. People cannot write pieces of paper for only social purposes, for example, by writing emails or messages, but also for academic and professional purposes in the form of reports, essays, summaries, government documents, newspapers, magazines etc. (Harmer, 2007, p. 80).

Writing is a crucial communicative skill in English as a first language (L1), second language (L2), and foreign language (EFL) teaching, learning, and testing instructions. It is one of the most challenging skills to be learned in EFL contexts (Jusun & Yunus, 2018, p. 470) due to its sophisticated cognitive nature, in which the writer has to perform different actions simultaneously, such as planning, organising, writing, revising, editing, and publishing (Weigle, 2002, p. 4) to produce coherent and accurate performance. In this respect, Suleiman (2000, p. 155) stresses the multidimensionality nature of this process in assessment practices and language instruction and development. To assess English as a second or a foreign language in the educational system, teachers should devote a special part to focus on learners' writing skills. The best way to assess students' writing competences is to ask them to write an academic sample because the ability to write helps teachers predict their learners' proficiency levels and academic achievements. Students' written productions reflect not only their degree of comprehension of structure and content but also their capacities to communicate their needs and thoughts using the target language (Kansopon, 2012, p. 86).

One way to measure L1, L2, or foreign language learners' writing abilities is to integrate essays in their exams and to ask candidates to perform coherent and meaningful academic samples and communicate effectively in the target language. Essay tests are one of the prominent evaluative methods, whose major aim is

Correspondence concerning this article should be addressed to:

E-mail: raniazribi@ymail.com
DOI: 10.32038/ltf.2021.04.03

to elicit learners' responses to writing prompts and to investigate their performances accordingly by taking into account the possible effects of different variables on the examinees' ability (Cumming, 1997, 2001; Hamp-Lyons, 1990, 2003; Weigle, 2002).

Despite their major importance in assessing learners' writing performances, academic essay tests require subjective scorers' judgments (Peterson, 2008, p. 72). In fact, raters may vary in assigning scores to the same test taker's performance by underestimating or overestimating his writing ability. Hence, testing students' writing skills is a complex cognitive task that requires the combination of different rating traits (rating scale, rating criteria, scores reliability and validity) (Hamp-Lyons & Henning, 1991, p. 364). This may lead to potential scores variations, distortions in ratings and different scoring patterns and reading styles that threaten the reliability and validity of scores in assessing EFL learners' writings.

Raters have different perceptions of what constitutes a good writing sample. What is appraised by one rater is downplayed by another. In justifying their scores' assignments, raters may overlook some mistakes while others may magnify them in measuring students' language skills. On that account, raters' potential scores' variability and divergent rating judgments can be due to various factors, namely scoring methods, essay features, and writing modes or task types (Cumming et al., 2002; Weigle, 2002; Schoonen, 2005).

Out of the various influential sources of scores' variation, this study focuses on both rater and task variables because raters who apply rating rubrics in language testing are important factors that influence scoring outcomes (Jeong, 2019, p. 2). Writing marks can also be affected by different tasks (Kim et al., 2017, p. 4). Hence, this paper casts light on two major variables that I deem to be influential in performance writing assessment settings: Raters and tasks. In this respect, Schoonen (2005) elucidates that both tasks and raters have a great impact on the scores' assignment task (p. 15). In terms of task types, the current study focused on narrative, and argumentative writing prompts, whose different rhetorical characteristics may have an impact on raters' judgments.

The major objectives of this study are as follow:

1. To investigate the interaction effects between raters' experience and narrative and argumentative tasks on raters' severity and internal consistency and the analytic rating scale functionality.
2. To diagnose the interaction effects between raters' experience and the two writing tasks on the raters' qualitative judgment process to show which aspects of language the novice rater group directed their attention to and which features attracted the experienced rater group's attention in assessing the two writing modes respectively.

The current study also addresses the following research questions:

a. Is there any interaction effect between rater groups and task types on raters' severity and internal consistency rates based on the analytic rating scale?
b. Is there any interaction effect between rater groups and task types on the analytic rating scale's performance?
c. Is there any interaction effect between rater groups and task types on the writing aspects experienced and novice raters attend to?

Literature Review

Academic writing is one of the prerequisite skills that learners need to fulfil different educational purposes, like sitting for national and international examinations and meeting specific communicative needs in English for Academic Purposes instruction (Zohaib et al., 2021, p. 3193). The development of writing skills is highly recommended in both L1, L2 and foreign language instruction depending on the learners' different learning interests, needs and objectives in various academic contexts.

Teachers value the importance of writing instruction based on diverse evidence-based and instructional practices, like conferencing, in order to develop learners' writing abilities. Special emphasis is put on additional instructional strategies, appropriate writing tasks, and potential adaptations for low-proficiency writers (Graham, 2019, p. 181). In fact, writing seems an indispensable subject of the academic syllabus in the educational system. At basic levels, the major focus of the schools' curriculum is on teaching writing skill to L1 learners (Weigle, 2002, p. 7) to help them process correct sentences. Hence, according to Kansopon (2012, p. 86), writing competence is an integral skill to help writers communicate effectively. Thus, the teaching of writing in a formal context requires an assessment phase.

To assess learners' writing skills, different tasks should be selected and adopted depending on the purpose of a test and its particular context of use. Hence, task-based assessment, including authentic academic tasks, help teachers activate and observe how the target language is used and interpret learners' language productions in real-life communicative contexts based on well-defined rating criteria (Giraldo, 2020, p. 212).

Judging candidates' performance constitutes a major concern in foreign writing measurements. Raters held different views about the applied rating scale. They even differed in their rating process, mainly when identifying salient text features, interpreting rating rubrics, and weighing the value of various task features to assign their final marks (Ghanbari & Barati, 2020, p. 9). Thus, raters exercise different reading strategies to come up with their final rating decisions on the quality of the candidates' writing performance. The interaction

of multiple rating factors may have an impact on making evaluative judgments of test takers' writing competences. Hamp-Lyons (1990, p. 82), for instance, propounds that different facets of variables, such as tasks, writers, scoring procedures, and raters interact with each other leading to a complicated network of effects that are difficult to be controlled. So, score variability has been seen as a source of measurement error that not only affects but also reduces the reliability of assessment outcomes and threatens the validity of essay scores (Huot, 2002).

Evaluating performances in the direct writing assessment field seems to be a challenging process as different text dimensions may interact with test-takers or raters' variables to affect the writing quality (Weigle, 2002 p. 65). Being two main variables in this study, raters and tasks. Although raters and tasks are two major variables in this study, the interaction between them remains questionable.

Several empirical research has pointed to the conceivable impacts of rater-task interaction on assessing learners' performances. In a study conducted by In'nami and Koizumi (2015), the effects of the interaction between different rating variables, including test takers, tasks, raters, and scoring criteria, should not be neglected in all designs. Considerable importance should be given to rater-task interaction effects in test design, development and validation science task, and task-related interaction effects were greater than the independent effects of tasks or raters on the assigned marks. (p.16). The interaction between the rater and other textual features, viz. the prompt topic, is further emphasised in Lumley's (2005) study. Raters should not be considered as just an interpreter of the writing sample. He may interact with other variables to affect the score assignment task and the overall essay judgment (p. 5).

Rater-task interactions can be further examined in Weigle's (1999) study. Thus, inexperienced, untrained scorers appeared harsher in marking EFL written comments on a graph than in assigning scores to independent topics. Consistent marks were, however, received from experienced raters in terms of their leniency and harshness patterns. Scores can vary due to either qualitative textual features or specific aspects related to the evaluation task (p. 146), namely task-rater interaction effects on the marks' assignment process. Hence, extensive variations in the scores assigned by raters to test takers' responses to different task types should be dealt with in order to ensure reliable judgments and fair marks (McNamara, 1996, p. 122). Interaction between different variables is thus another salient factor in examining raters' scores and judgments (Barkaoui, 2008; Erdosy, 2000).

The role of both raters and tasks in the testing process corroborates with Schoonen's (2005) view that the effects of tasks and raters are based on what has to be marked in the evaluated performance and how it has to be judged (scoring procedures) (p. 5). On that account, differences in task characteristics can, directly and indirectly, affect raters' rating behaviours of EFL learners' responses to writing tests.

Clearly, based on the literature covering the factors that may affect the raters' assessments of learners' writing skills, special attention should be given to the scorers' rating behaviours in addition to other scores' variability sources, like raters and tasks in the language testing context. Identifying the sources of variations in the scores assignment task helps explain variability in measuring candidates' writing skills in this study. Two potential sources of variance, raters and prompt types are investigated in this EFL writing assessment study.

Methodology

Research Design Overview

This study addresses two major questions, namely "how" did rater' groups as an independent variable assess the learners' responses to two distinct prompts based on the same rating criteria and "why" did they assign such similar or different scores to the same candidates' narrative and argumentative performances. These questions help me to diagnose any possible interaction effect between raters and tasks.

Emphasis was put on two independent variables: raters' groups and the two different writing tasks to examine rater-task interaction effects by extracting any possible relationship between the two independent variables and their potential effects on the assessment process.

A correlational research design was selected in order to determine any relationship between rater groups and tasks after assessing third-year English Students' writing performances. Based on the statistical program FACETS, correlation coefficients reflected the degree of the interaction between the two variables. The interpretation of results is thus based on providing empirical evidence for any possible rater-task interaction after the assessment process. In fact, I intended to observe and to probe whether the two groups of raters differed in their severity and consistency levels after assessing their test takers' narrative and argumentative essays. I also tried to show which aspects of language the novice rater group directed their attention to and which features attracted the experienced rater group's attention in assessing the two writing modes, respectively.

Participants

A panel of thirty EFL learners participated in this study. As a representative sample of a large population, these students were enrolled in the third year English class level. They were mostly females with a mean age of 22. They were undergraduate third-year English students who have been specialised in English as a foreign

language for three years at the tertiary level and whose proficiency levels vary. All the examinees were non-native speakers of English and students in the English department at the Faculty of Arts and Humanities.

Moreover, a total sample of fifty writing teachers of English as a foreign language took part in this phase. They represented a mixed sample of male and female raters with an average age of 45 and belonged to different L1 backgrounds. Their first language is Arabic, while English is their dominant work language in tertiary education in different Tunisian universities. The two rater groups were mainly distinguished in this thesis based on their rating experience in assessing EFL learners' writing skills. The experienced rater group served as raters for more than ten years. The rating experience for the novice rater group, however, ranged from one to three years in judging learners' performances.

Procedures

To meet the purpose of this study, my aim was to examine the way raters assess and score EFL third year learners' narrative and argumentative essays by focusing on any possible effect of two different tasks, especially various aspects of writing tasks on rater groups' scores, which may lead to rater-task interaction. It was hypothesised that different writing modes affect raters' scoring behaviours of learners' performances, and there may be a rater-task interaction effect in measuring test takers' narrative and argumentative essays.

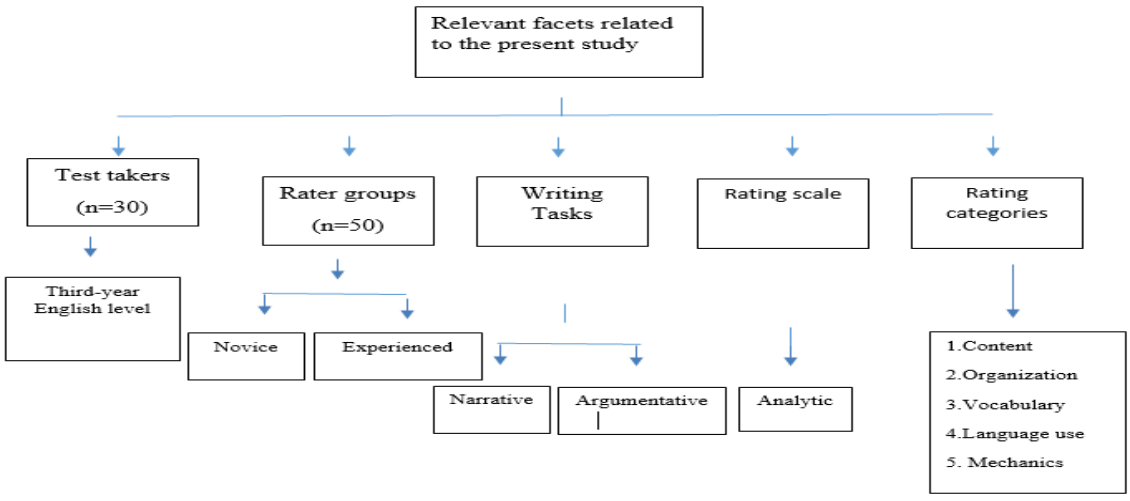
In this study, each test taker responded to two different narrative and argumentative tasks to come up finally with a total of sixty essays. Both novice and experienced tertiary level teachers of English made judgments on this set of test takers' narrative and argumentative writing samples. Students' compositions were assessed analytically by taking into account five criteria, namely content, vocabulary, language use, organisation, and mechanics.

Since this study has embarked on examining raters' scoring patterns of EFL learners' writing performances in both narrative and argumentative discourse modes, a mixed-methods triangulation design of both quantitative and qualitative approaches was applied to gather and report data about the raters' decision-making while rating under-graduated learners' essays.

To gather quantitative data, raters wrote the scores that they assigned to the sixty narrative and argumentative compositions in the first part of the analytic score report forms. These scores were then analysed based on two statistical programs: SPSS and FACETS (version 3.80.0). The latter permits researchers to add the necessary facets of variables, such as rater groups, tasks, students, rating scale, rating criteria, and so on, depending on the aim of the research. In this respect, Karlin (2018) claims that "if it can be agreed upon that the Rasch measurement model provides better and more accurate information than raw scores, then the only excuse for not using the Rasch measurement model is that the process might be too complicated". (pp.98-99). Prior to the analysis and interpretation of raters' judgments, the facets used in this study were coded. The following figure presents the relevant facets related to this thesis in the data collection phase.

Figure 1

A Schema of the Relevant Facets of Assessment to this Study



This figure illustrates the major facets used in the study. Thirty test-takers responded to two different academic narrative and argumentative tasks. Their written performances were then assessed by two rater groups (experienced and novice rater groups) based on an analytic rating scale which contains five different scoring criteria, namely content, organisation, vocabulary, language use, and mechanics.

Moreover, the second part of report writing forms and think-aloud protocols were also examined in order to collect qualitative data about raters' judgments of EFL test takers' writing responses to two different narrative and argumentative prompts. Raters' reading and assessing strategies were thus elicited based on explaining their rating patterns and verbal reports during the evaluation process. Mixed methods are useful in this study in order to draw some information about which aspects of the narrative and argumentative task modes attract the attention of experienced and inexperienced raters in measuring students' written productions. This part helps clarify the rater-task interaction by taking into account language criteria that experienced, or novice raters may focus on at the expense of others during the judgment process.

Results

This section reports findings concerning the interactions between rater and task facets based on the analytic ratings. It examines the effects of the rater group-task interaction on raters' severity and consistency levels, scale functioning, and the writing aspects that attracted the raters' attention across tasks.

Interaction Analysis

Interaction analysis, also known as bias analysis, examines each pair of facets and reports bias t-statistics estimated in logits, showing the significant and insignificant biased pattern of analytic scores. The bias analysis of FACETS provided information about the unexpected raters' reactions to particular assessment aspects with respect to their characteristics. As Wigglesworth (1993) put it, bias analysis investigates "the characteristics of each particular rater with respect to each specific task or criterion in the test (or) investigates whether particular raters rate more or less harshly or leniently on a particular criterion of the test" (p. 307). These statistical bias estimates test the null hypothesis that no bias, except for measurement error, can be found in the raters' scores to examinees' responses to two writing prompts. A two-way interaction analysis was used in this study to find out sources of variability in test scores. Interaction analysis is based mainly on examining tables extracted from FACETS output. The bias size is estimated in logits. T-scores dealing with scores' variation reflect significant and insignificant bias. T-values that fall between this range: $+ 2 <$ t-scores < -2 indicate significantly biased interactions. Values higher than + 2 indicate that raters are harsher in rating that particular assessment aspect, whereas values lower than -2 indicate that raters are more lenient in rating a specific testing criterion. (Barkaoui, 2008, p. 26).

Interaction Effects on Scoring Learners' Compositions (quantitative analysis)

After clarifying the issue of rater-task interaction, it is important to diagnose the effect of the interaction between rater groups and task types on raters' severity and consistency levels and the analytic scale's functioning based on the analytic scores awarded by novice and expert rater groups to the test takers' responses to narrative and argumentative writing modes.

Rater Severity

The rater group-by-task type interaction highlighted some differences in terms of rater severity estimates based on the analytic marks assigned by both rater groups to the narrative and argumentative productions. According to table 1, the experienced rater group was harsher than his novice counterpart in assessing both types of essays. Test takers' narrative performance received lower scores compared to the argumentative writings. Thus, the rater groups' tendency towards leniency was highlighted in measuring the examinees' argumentative skills. The severity range estimates in scoring narrative essays were 4.92 for the experienced rater group and 3.99 for the novice rater group. The experienced rater group reached a severity range of 3.85 logits, whereas the novice rater group spanned the range of 3.61 logits in marking examinees' argumentative essays. Raters' average severity exhibited larger variability in rating the narrative writings (.93 logits) than the argumentative writings (.24 logits). As the table below indicates, the separation index and reliability of separation values were higher in the narrative task than in the argumentative one for the two rater groups.

Each rater group displayed different average severity levels across task types. A tendency towards leniency appeared in the scores assigned by the novice rater group (mean severity: 3.80 logits) to the candidates' sixty written samples, compared to the experienced rater group (mean severity: 4.36 logits). Furthermore, the separation index (H) and the reliability of separation statistics were significantly higher for the experienced group (H = 12.63 and 10.98; reliability of separation = .99 and .98 for both tasks) compared to the novice group (H = 11.61 and 10.86; reliability of separation = .98 and .97 for the two modes).

Table 1
Summary of Rater Measurement Report across Rater Group and Discourse Modes (Based on the Analytic Scoring Procedure)

		Narrative Task	Argumentative Task
Experienced Rater Group	Mean	.13	.13
	SD	.00	.01
	Min	-1.80	-2.33
	Max	3.12	1.52
	Separation Index (H)	12.63	10.98
	Reliability of separation	.99	.98
	Observed Agreement %	38.4%	37.7%
Novice Rater Group	Mean	.13	.14
	SD	.00	.01
	Min	-2.01	-1.87
	Max	1.98	1.74
	Separation Index (H)	11.61	10.86
	Reliability of separation	.98	.97
	Observed Agreement %	37.3%	39%

The relationship between rater severity estimates for each rater group across writing modes in measuring test takers' writing skills analytically was examined based on the Pearson correlation statistical test. Some differences appeared between the two rater groups, as the correlation values were slightly higher for the novice rater group ($r = .34$) than for the experienced rater group ($r = .29$). To examine the way raters' severity estimates were ordered in each group, a Wilcoxon Signed-Rank test was applied in this analytic part. Statistical outcomes reflected significant differences in ranking novice ($z = -10.03$ $p = .00$) and experienced raters ($z = -8.7$ $p = .00$) across the narrative and argumentative prompts, as their p-values were .00.

Concerning the experienced rater group, sixteen raters displayed more severity in assigning scores to the narrative essays, whereas nine raters exercised more harshness in rating the argumentative essays. There seemed to be also some variability in the severity range of the novice rater group. Seven inexperienced raters were more severe in grading narrative samples, while eighteen novices were harsher in assessing argumentative samples.

Rater Internal Consistency

Interaction effects on raters' internal consistency were scrutinised based on FACETS outcomes across rater groups, and writing prompts. A perfect mean infit statistical value of 1.00 was perceived in the analytic scores assigned by the experienced rater group to the candidates' narrative and argumentative essays. A slightly lower infit mean value (.99) was detected based on the marks given by the novice rater group than the experienced one to the same performance. Table 2 reports the frequencies of novice and experienced raters, whose analytic scores fell in an acceptable fit, overfit, or misfit ranges, in testing the same examinees' narrative and argumentative productions.

Table 1
Frequencies of Rater Fit Statistics Across Rater Groups and Discourse Modes (Based on the Analytic Scoring Procedure)

	Narrative Mode		Argumentative Mode	
Fit Range	**Experienced Group**	**Novice Group**	**Experienced Group**	**Novice Group**
OverfitF < 0.70	4 (16%)	3 (12%)	3 (12%)	2 (8%)
Acceptable Fit 0.70 < F < 1.30	19 (76%)	16 (64%)	21 (84%)	19 (76%)
MisfitF > 1.30	2 (8%)	6 (24%)	1 (4%)	4 (16%)

Raters from both groups displayed higher acceptable consistency rates in measuring test takers' argumentative essays than narrative ones. The number of raters with overfitting and misfit decreased in assessing argumentative essays. The marking of argumentative writings resulted in some obvious differences in rater groups' consistency rates, as more overfitting experienced raters appeared in the rating process compared to the novice rater group. About 24% of the novice rater group exercised more misfit in rating narrative performance than did the experienced rater group.

More experienced raters (76% in the narrative mode and 84% in the argumentative mode) showed acceptable infit in their scoring behaviours than did the novice rater group (64% in the narrative mode and 76% in the argumentative mode). Out of the twenty-five novice raters, six novices showed misfit in measuring learners' narrative productions, and four novices exhibited misfit in grading argumentative writings. Slight differences were perceived in overfitting raters. Four and three experienced raters demonstrated more overfit in

assigning scores to both narrative and argumentative essays respectively compared to the novice rater group (three overfittings raters in the narrative mode and two with overfit in the argumentative mode). The effects of the interaction between rater groups and task types on raters' internal consistency in grading their test takers' performances analytically were statistically significant, as their p-values were .00.

Analytic Scale Performance

FACETS statistical output, especially examinees' ability measures and outfit values, was analysed to examine the analytic rating scale performance in scoring test takers' narrative and argumentative writings across rater groups and writing modes. According to table 3, examinees' ability estimates increased as the analytic rating levels advanced. They ranged from -1.65 in the lowest category 1 to 1.58 in the highest category 4 for the experienced rater group and from -1.69 in category 1 to 1.84 in category 4 for the novice rater group in marking the same candidates' narrative skills.

Table 2

Average Measures and Outfit Indices for a Four-Category Analytic Rating Scale across Rater Groups and Task Modes

		Experienced Rater Group		Novice Rater Group	
	Category	**Examinees' measures**	**Outfit MS**	**Examinees' measures**	**Outfit MS**
Narrative Mode	1	-1.65	1.1	-1.69	1.2
	2	-.72	.9	-.49	.9
	3	.48	1.0	.81	1.1
	4	1.58	1.0	1.84	1.0
Argumentative Mode	1	-1.76	1.0	-1.90	1.2
	2	-.51	1.1	-.66	1.0
	3	.63	.9	.97	1.1
	4	1.94	1.0	2.66	.9

Some differences appeared in the test takers' average scores assigned to the argumentative samples across rater groups from the lowest category 1(-1.76 logits for the experienced rater group whereas -1.90 for the novice rater group) to the highest category 4 (1.94 logits for the experienced group while 2.66 for the novice group). Thus, each analytic rating category was used by both experienced and novice rater groups in testing narrative and argumentative essays. Examinees writing ability measures progressed monotonically with the four analytic rating categories, from category 1 standing for the students' lowest writing ability to category 4 representing the students' highest writing ability estimates. The analytic rating scale, with its four levels, functioned adequately across rater groups and task types.

The functionality of the analytic rating scale was further examined based on outfit mean square values. As table 3 indicates, the experienced rater group attained the expected outfit value of 1.00 in categories 3 and 4, which reflected similarity between the expected and the observed examinees' ability measures in rating narrative essays. The novice rater group, on the other hand, reached the outfit value of 1.00 only in category 4 in assessing the same narrative samples. The outfit values were between .9 and 1.1 for the experienced group and from .9 to 1.2 for the novice group.

In their argumentative scores' assignment task, the outfit mean square value was 1.00 for the experienced rater group in three different categories, 1, 2, and 4. An expected outfit value of 1.00 was noticed in just the second analytic rating category for the novice rater group. Apart from detecting expected outfit value of 1.00 in the scores given by both rater groups, the experienced group reached an outfit value of .9 in the third category, whereas the outfit values were .9 in the fourth category, 1.1 in the third category, and 1.2 in the first category of the analytic rating scale. To conclude, the four analytic rating categories functioned as expected by FACETS across rater groups and task modes.

Interaction Effects on Qualitative Judgments across Rater Groups and Writing Modes

The previous section examined the impact of the rater group-by-writing mode interaction on raters' scoring severity and internal consistency and the scales' performance based on the quantitative analysis of the analytic marks assigned by raters to the same set of narrative and argumentative compositions. The effect of the interaction between the same variables on the writing aspects attended to by each rater group to each writing mode is further highlighted in this qualitative part. It shows which aspects of language attracted novice and experienced rater groups in assessing narrative and argumentative writing performance by interpreting analytic comments and scores' explanations and analysing think-aloud protocols across rater groups and task types.

Raters scores' explanations

This qualitative part aims at diagnosing the rating patterns and writing aspects across not only rater groups but also writing modes in the analytic rating scale. During their assessment process, both rater groups were aware of the narrative and argumentative genre stipulations and task-specific requirements. However, some differences appeared in their decision-making behaviours. A major question that could be at the heart of this qualitative analysis part is which aspects of writing were deemed more important than others for each rater group in each writing task.

Some differences appeared in rater groups' testing of the same test takers' narrative and argumentative writing abilities for each essay type based on the analytic rating scale. Other writing aspects, vocabulary, and language use were reported more frequently by the experienced rater group, while both content and language use were reported more frequently by the novice rater group in evaluating narrative essays. In scoring argumentative essays, the experienced rater group made more comments on organisation and vocabulary than did the novice raters, who made more comments on vocabulary, organisation, and language use. In terms of other writing aspects, the experienced raters relied more on their overall impressions, narrative style, tenses, and originality than did the inexperienced raters, which focused on narrative task completion.

A special focus on formal style, tense and essay length was highlighted by the experienced rater group in measuring examinees' argumentative skills. The novice raters, however, reported parallelism and English rules more frequently than did the former group in judging the same argumentative essays. It seems that rater groups were influenced by specific task characteristics in judging different writing modes. Raters may have some stylistic, syntactic, or semantic preferences while measuring learners' argumentative essays than narrative ones.

The experienced raters made more comments on vocabulary based on appropriate lexical words and form choices than did the novices, whose major attention was directed to the way examinees used similes and descriptions to perform narrative essays. The experienced group paid more heed to coherence, as compared to the novice rater group by stressing the influence of inappropriate vocabulary choice and language error gravity on the narrative essays' coherence and comprehension. Concerning the argumentative essays, the experienced rater group was prone to focus more on balance between thesis and anti-thesis paragraphs than did the novice raters, who paid more attention to the specific components of the introduction, namely the motivator, background information, thesis statement, and blueprint and the recommendations in the concluding paragraph.

The novices provided more explanations on language use by mentioning word order and placement, idiomatic expressions, phrasal verbs, long ambiguous sentences, and subject-verb agreement than did the experienced raters, who referred to grammatical accuracy in judging argumentative compositions. The experienced raters' greater attendance to linguistic and grammatical accuracy in their rating process than that of the novice rater group can be due to their teaching, rating and training experiences (Cumming et al., 2001; Erdosy, 2004; McNamara, 1996; Song & Caruso, 1996). The fact that the novice rater group focused on the specific components of introduction and conclusion in testing both writing genres can be due to their previous learning patterns. They relied on what they learnt concerning academic writing modes.

Think Aloud Protocol Analysis

This section explores and compares four experienced and four novice raters' judgments of the first test taker's narrative and argumentative performance based on the think-aloud protocol method. A random set of four experienced raters and four novices were asked to evaluate the first examinee's narrative and argumentative essay based on the same analytic rubric. Both rater groups' verbal reports were recorded while assessing the writing samples. Their verbal reports were then transcribed and coded to analyse and interpret raters' decision-making behaviours.

Think aloud analysis of all participants was based on the coding scheme of Cumming et al. (2002) as it focuses on language and its rhetorical and ideational aspects along with interpretation and judgment strategies, which meets the main objective of this study. In what follows, tables 4and 5 lists the percentages of rating styles and writing aspects mentioned in the think-aloud protocols based on each separate category and main strategy across rater groups and task types. Some differences were perceived in comparing think-aloud categories and strategies across the two writing modes. The two rater groups focused more on self-monitoring focus and interpretation strategies in assessing the narrative writing mode and more on judgment strategies in testing the argumentative writing mode. Recall that self-monitoring strategies consider reading and interpreting compositions, revising rating criteria, summarising and comparing them with other essays, and articulating general impressions. Rhetorical and ideational strategies focus on ambiguous phrases, rhetorical structures, topic development, task completion, relevance, coherence, and organisations. Finally, language strategies deal with language errors, essays' comprehensibility, language, as well as mechanics and lexical, morphological and syntactic aspects.

Looking for differences in reading styles and rating behaviours between the two groups of raters and between the two writing modes, some variations were also detected in which categories and strategies were more reported than others by each rater group in each writing mode. On the one hand, the experienced rater group reported more rhetorical and ideational focus and judgment strategies than did the novices in rating the

narrative and argumentative modes. The novices, on the other hand, reported more language focus in terms of both judgment and interpretation strategies than did the experienced raters in assessing the two types of essays. As the tables show, the experienced raters referred more to self-monitoring and interpretation strategies than did the novices who pointed more to self-monitoring and judgment strategies in measuring the narrative and argumentative samples. However, rhetorical and ideational judgment strategies were much more displayed by the experienced raters than by the novice raters in judging both essays. The novices thus devoted more attention to the interpretation strategies within the rhetorical and ideational category compared to their experienced counterparts in the two writing tasks. These differences in the decision-making behaviours across rater groups reflected the qualitative impressions that the experienced raters, as a group, balanced their attention to rhetorical and ideational issues and to judgment strategies fairly while scoring both writing modes. On the contrary, the novice rater group made more comments related to language focus and self-monitoring judgment strategies than they did to the rhetorical and ideational category in their evaluations of the same narrative and argumentative essays.

Apart from diagnosing the main think-aloud protocol categories and strategies, some differences could be highlighted in terms of rater groups' focus on sub-categories in each writing mode. The experienced raters reported more sub-categories in articulating their verbal reports to assess the test taker's narrative essay than did the novice raters. In this vein, they tended to devote more attention than the novice raters did to reading or interpreting the essay prompt, scanning the whole composition, articulating general impressions, discerning the rhetorical structure, assessing coherence, assessing text organisation, assessing style, register, or genre, assessing fluency, assessing lexis, considering spelling or punctuation, and rating language overall. In contrast, the novices tended to pay more attention than the experienced raters did to revising their own criteria, articulating or revising scoring, interpreting ambiguous or unclear phrases, assessing text organisation, and considering syntax or morphology in assessing the same narrative genre. The experienced raters' verbal protocols reflected the different sub-categories that raters relied on testing the argumentative essay. They made more comments related to comparing this essay with other compositions, summarising, distinguishing, or judging collectively, discerning rhetorical structure, summarising ideas or propositions, assessing text organisation, and considering spelling or punctuation. Fewer verbal reports were extracted from the novices' protocols compared to the experienced raters while measuring the argumentative sample. It was obvious that raters with extensive rating experience tended to verbalise their decisions by producing longer and more detailed think-aloud protocols than did the inexperienced raters.

The novices reported the following sub-categories more frequently, namely, assess text organisation, assess the quantity of total written productions, assess the gravity of errors, assess error frequency, consider syntax or morphology, and consider spelling or punctuation.

Table 3

Experienced Raters' Think Aloud Protocols by Writing Mode

	Narrative Mode		Argumentative Mode	
Think-Aloud Coding Scheme for Experienced Rater Group	**Number of times used**	**Percentages**	**Number of times used**	**Percentages**
Self-monitoring-Interpretation	24	3.5%	9	2.22%
Self-monitoring-Judgment	32	4.76%	8	1.97%
Rhetorical-Interpretation	22	3.27%	15	3.70%
Rhetorical-Judgment	72	10.71%	60	14.81%
Language Interpretation	10	1.48%	3	0.74%
Language Judgment	64	9.52%	40	9.87%
Self-monitoring	56	8.33%	17	4.19%
Rhetorical and Ideational	94	13.98%	75	18.51%
Language	74	11.01%	43	10.61%
Interpretation	56	8.33%	27	6.66%
Judgment	168	25%	108	26.66%

Table 4
Novice Raters' Think Aloud Protocols by Writing Mode

	Narrative Mode		Argumentative Mode	
Think-Aloud Coding Scheme for Novice Rater Group	**Number of times used**	**Percentages**	**Number of times used**	**Percentages**
Self-monitoring-Interpretation	2	1.90%	1	0.92%
Self-monitoring-Judgment	9	8.57%	3	2.77%
Rhetorical-Interpretation	8	7.61%	2	1.85%
Rhetorical-Judgment	4	3.80%	4	3.70%
Language Interpretation	3	2.85%	3	2.77%
Language Judgment	9	8.57%	24	22.22%
Self-monitoring	11	10.47%	4	3.70%
Rhetorical and Ideational	12	11.42%	6	5.55%
Language	12	11.42%	27	25%
Interpretation	13	12.38%	6	5.55%
Judgment	22	20.95%	28	25.92%

These results suggested that the novice raters reported particular language aspects, such as error gravity and frequency, syntax, morphology, spelling, punctuation, organisation, and lexis in their think-aloud protocols during their argumentative and narrative assessments, which was not the case for the experienced raters, who made more sophisticated comments on ideational and rhetorical focus during their progressive decision making while reading. This can be traced back to the novices' lack of testing and rating experience in assessing the same test taker's essays. They were prone to focus on some basic aspects of language by taking into account error frequency and gravity. Due to their teaching and rating experience, the experienced rater group focused on rhetorical structure and ideational aspects with their judgment strategies by referring to style, coherence, register, in particular, and to text organisation in general.

Based on what was mentioned earlier, the novice raters tended to revise their scoring criteria and articulate or revise scoring more frequently than did the experienced raters, who concentrated more often on scanning the whole composition and reading and interpreting essay prompts while testing the narrative essay. This can be explained by the fact that the novices were more dependent on the rating rubrics and their specific descriptions than were the experienced raters. The novices invested their time in revising scales' criteria to articulate scorings, which may reflect their relative lack of rating experience compared to their experienced counterparts. The latter, however, devoted more time to scanning, reading and assessing composition. Hence, comparable think-aloud protocols differed qualitatively not only by main category and strategy but also by their sub-categories across rater groups and writing modes. Rater groups appeared to display different ranges of verbal reports while assessing narrative and argumentative writings.

Discussion

FACETS quantitative analysis provided a more detailed picture of the rater-task interaction effects on test takers' writing ability estimates and raters' severity and internal consistency measures. The outcomes of this study revealed some statistically significant differences in the scores' assignment process across rater groups and task types. Different tasks seemed to have a significant effect not only on test takers' writing ability estimates ($p = .00$) but also on raters' severity and internal consistency. Thus, examinees were ranked differently across rater groups and task types analytically. As the outcomes of this study showed, task effects were consistently wider than rater effects. This concurs with Brennan and Johnson's (1995) view that different studies came up with the conclusion that the person-by-task interaction effects were larger than person-by-rater interactions (p.9).

Despite their overall leniency, novice raters in this piece of research exhibited slight differences across tasks. However, the majority of expert raters displayed more severity in assessing narrative essays than argumentative essays. Experienced severity tendency in scoring narratives can be due to their previous rating experience of narrative writings. Their engagement with the prompt type can be related to their rating expectations and prior marking backgrounds. This can be explained by the fact that some expert raters with creative rating experience, for instance, exerted harshness in assessing narrative essays as they expected to receive original creative narrative essays from students rather than a series of chronological events that were narrated in a simple non-creative way, which according to them did not meet the narrative mode requirements.

More overfitting experienced raters, and misfitting novice raters also appeared in assessing narrative and argumentative tasks. But, it seems that fewer misfit and overfit rates appeared in the argumentative mode

compared to the narrative mode. As these biased interactions between rater groups and task types were significant ($p < .005$), these differences across rater groups and task types were more likely to be systematic rather than accidental (by chance). These findings suggest that it is crucial to take into account interaction effects of different facets or variables in writing assessments. The importance of interaction effects was further advocated by some researchers, such as Schaefer (2008), who argued that the idea of searching for unexpected interactions among rater judgments and test takers' performance or other facets in the analysis is central to bias analysis. It can identify patterns in ratings unique to individual raters or across raters and whether these patterns or combinations of facet interactions affect the estimation of performance (p. 467).

In assessing narrative and argumentative writings, experienced and novice rater groups attended to different writing aspects and distinct sub-rating criteria. This can be explained by their rhetorical, syntactic, and stylistic preferences and strategies associated with a particular writing task. Both rater groups valued the importance of each genre requirement, but they differed in interpreting and judging narrative and argumentative samples. The experienced raters directed their attention to more writing aspects as well as some rating criteria than the novices in testing narrative essays.

On the contrary, the inexperienced rater group concentrated on different features under the language use criterion, vocabulary and organisation without referring to other writing aspects that are not mentioned in the rating scales. This seems to be due to the experienced raters' previous rating experiences, past training practices, reading and rating schemes and styles and familiarity with students' possible writing deficiencies. The experienced raters' greater attendance to linguistic and grammatical accuracy in their rating process than that of the novice rater group can be due to their teaching, rating and training experiences (Cumming et al., 2001; Erdosy, 2004; McNamara, 1996; Song & Caruso, 1996).

Some differences also emerged in the verbal reports produced by novice and experienced rater groups in judging narrative and argumentative writing modes. The expert rater group concentrated on rhetorical and ideational focus together with judgment strategies, while the novice rater group reported language focus and self-monitoring judgment strategies when rating the sixty performances. The former group spent more time evaluating their candidates' writings and gave more sophisticated comments than did the inexperienced rater group. This result was in line with other past studies (Cumming, 1990; Barkaoui, 2010).

Apart from the divergences in the think-aloud protocols' main categories, some variation can be detected in the rater groups' use of the sub-categories. The experienced raters resorted to scanning and interpreting essays while the novices re-read the rating criteria and revised scoring while producing their protocols. A possible reason for this finding is the experienced raters' confidence in judging learners writing skills critically. The novices, on the contrary, focused on surface structures by detecting language errors frequency and gravity due to their lack of rating experience, which may develop their cognitive processes and mental decisions in the writing assessment field.

Overall, then, some differences appeared in the scoring behaviours of experienced and novice rater groups in measuring learners' narrative and argumentative performances. Despite their severity levels, the experienced raters were more consistent in their scores' assignment task compared to the novices. Moreover, experienced raters assessed candidates based on their critical thinking strategies within a short period of time, which was not the case for the novice raters. Rating the sixty essays was a time-consuming task for the inexperienced raters, who referred much time to the analytic rating scale. Indeed, the notion of expertise in the language assessment field exists in theory as well as in practice. The differences between rater groups extended to their perception of task difficulty, especially the specific rhetorical tasks, and stylistic and syntactic preferences related to genre requirements.

This idea of experience in assessment can help in a number of ways. Novice raters can benefit from experienced raters by attending training sessions about how to use rating scales and assess test takers' writing skills in a critical way. This helps them to measure their candidates' performances more confidently than before by heeding their attention to deep language patterns rather than referring to surface-sentence problems.

As writing assessment is a time-consuming task, novice raters need to gain knowledge from their experienced counterparts about how to manage to evaluate their test takers' writing proficiency reliably and consistently without much reference to the rating scales. A possible collaboration between the inexperienced and experienced raters could be suggested to help the novices to internalise different rating schemes, patterns, strategies, and styles to develop their cognitive process.

Based on the major findings of the present study, it would seem that the narrative and argumentative tasks did indeed differ in difficulty, and a prompt effect might exist in some situations. Due to the different writing prompt requirements, the novice rater group should use specific task-based rating scales designed to assess each particular genre (narrative, argumentative) to help them internalise specific narrative and argumentative rating criteria, stylistic and syntactic writing features and task characteristics. Thanks to their prior teaching and rating experience, the experienced raters may guide and help the novices in assessing learners' language abilities accurately. Thus, a call for cooperation and expertise-sharing between novices and experienced raters is deemed important in order to bridge the gap between potential variability in the writing assessment field.

Conclusion

The effects of the interaction between rater groups and task types on scoring learners' essays were significant, suggesting some differences across raters and tasks. In the narrative mode, the experienced raters pinpointed other writing aspects, viz. authenticity, originality, overall impression, and narrative style, vocabulary (word/form choices), and language use (errors gravity) criteria. In the argumentative mode, however, they focused on the organisation (paragraphing and structure), vocabulary, other writing aspects (format style, essay length, tense), and grammatical accuracy. The novice rater group, on the other hand, highlighted content (descriptions), language use, and task completion (other writing aspects) in judging the narrative writings. More comments were given by the novices to the argumentative essays, in which they concentrated on vocabulary, organisation (introduction and conclusion components), and language use by taking into account English rules, word order, idiomatic expressions, sentence constructions, and verb forms. Another major difference can be detected in the think-aloud protocols produced by rater groups. The experienced rater group generated more verbal reports and used more thinking aloud strategies than the novice group in both task types. This can be due to the effect of their rating experience on judging and scoring learners' performances.

The findings of this research have some practical implications. These immutable differences may require the use of task-based rubrics related to the specificities and characteristics of each writing genre, one for the narrative and one for the argumentative. It is also proposed to apply analytic rubrics in performance assessments as they provide "detailed data to researchers about the characteristics of texts and the value raters ascribe to texts and text facets" (Hamp-Lyons, 1995, p. 761). The rating differences among rater groups suggest more collaboration between both rater groups, in which raters exchange and benefit from their formal and informal feedback that might play a key role in moderating raters' behaviours. It is also crucial to support the practicality of applying think-aloud protocol methods in future writing assessment studies by investigating, for example, the effects of thinking aloud, including various verbal reports, on judging learners' performances based on FACETS or any other statistical measures.

References

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes* [Unpublished doctoral dissertation]. University of Toronto, Toronto, Canada.

Barkaoui, K. (2010). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, Cross-sectional study. *TESOL Quarterly, 44*(1), 31-57. 10.2307/27785069

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice*, *14*(4), 9-12. https://doi.org/10.1111/j.1745-3992.1995.tb00882.x

Caswell, R., & Mahler, B. (2004). *Strategies for teaching writing*. United States of America: ASCD (Association for Supervision and Curriculum Development).

Cumming, A. H., Kantor. R., & Powers, D. E. (2001). Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework. *TOFEL Monograph Series*, MS-22. Princeton, NJ: ETS.

Cumming, A. (1990). Expertise in evaluating second language composition. *Language Testing*, *7*(1), 31-51. https://doi.org/10.1177/026553229000700104

Cumming, A. (1997). The testing of writing in a second language. In C. Clapham & D. Corson (Eds.). *Encyclopedia of language and education, vol.7* (pp. 51-63). Dordrecht, The Netherlands: Kluwer.

Cumming, A. (2001). The difficulty of standards, for example in L2 writing. In T. Silva & P. Matsuda. (Eds.) *On Second Language Writing* (pp. 209-229). Mahwah, NJ: Lawrence Erlbaum.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*(1), 67-96. https://doi.org/10.1111/1540-4781.00137

Erdosy, M. U. (2000). *Exploring the establishment of scoring criteria for writing ability in a second language, the influence of background factors on variability in the decision-making processes of four experienced raters of ESL compositions* [Doctoral dissertation]. National Library of Canada.

Erdosy, M. U. (2003). Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions. *TOEFL Research Report* N.70. 10.1002/J.2333-8504.2003.TB01909.X

Giraldo, F. (2020). Task-Based Language Assessment: Implications for the Language Classroom. *GiST-Education and Learning Research Journal, 21*, 209-224. https://doi.org/10.26817/16925777.828

Graham, S. (2019). Changing how writing is taught. *Review of Research in Education, 43*(1), 277-303. https://doi.org/10.3102/0091732X18821125

Hamp-Lyons, L. & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning, 41*(3), 337-373. https://doi.org/10.1111/j.1467-1770.1991.tb00610.x

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In Kroll, B. (Ed.). *Second language writing: Research insights for the classroom* (pp. 69-87). Cambridge University Press.
Hamp-Lyons, L. (1995). Rating Non-native writing: The trouble with holistic scoring. *TESOL Quarterly, 29*(4), 759-762. 10.2307/3588173
Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In B. Kroll. (Ed.). Exploring the dynamics of second language writing (pp.162-189). Cambridge University Press. http://hdl.handle.net/10397/69136
Harmer, J. (2007). *The practice of English language teaching*. Harlow: Pearson Longman.
Huot, B. (2002). *(Re) Articulating writing assessment: Writing assessment for teaching and learning.* Utah State University Press. https://digitalcommons.usu.edu/usupress_pubs/137
In'nami, Y., & Koizumi, R. (2015). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing, 33*(3), 341-366. https://doi.org/10.1177/0265532215587390
Jeong, H. (2019). Writing scale effects on raters: an exploratory study. *Language Testing in Asia, 9*(20), 1-19. https://doi.org/10.1186/s40468-019-0097-4
Jusun, K. D. & Yunus, M. M. (2018). The effectiveness of using sentence makers in improving writing performance among pupils in Lubok Antu rural schools. *International Conference on Education (ICE2) Education and Innovation in Science in the Digital Era*, 469-475.
Kansopon, V. (2012). An Investigation of the writing test used at the institute of international studies. *Language Testing in Asia*, *2*(4), 86-100. https://doi.org/10.1186/2229-0443-2-4-86
Karlin, O., & Karlin, S. (2018). Making Better Tests with the Rasch Measurement Model. *InSight: A Journal of Scholarly Teaching*, *13*, 76-100. 10.46504/14201805ka
Kim, Y. S. G., Schatschneider, C., Wanzek, J., Gatlin, B., & Al Otaiba, S. (2017). Writing evaluation: rater and task effects on the reliability of writing scores for children in Grades 3 and 4. *Reading and writing, 30*(6), 1-25. 10.1007/s11145-017-9724-6
Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Cambridge University Press.
McNamara, T. F. (1996). *Measuring second language performance*. New York: Longman.
Ghanbari, N., & Barati, H. (2020). Development and validation of a rating scale for Iranian EFL academic writing assessment: a mixed-methods study. *Language Testing in Asia, 10*(17), 1-21. https://doi.org/10.1186/s40468-020-00112-3
Peterson, S. S. (2008). *Writing across the curriculum: All teachers teach writing*. Portage & Main Press.
Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493. https://doi.org/10.1177/0265532208094273
Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22*(1), 1-30. https://doi.org/10.1191/0265532205lt295oa
Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking, and ESL students? *Journal of Second Language Writing, 5*(2), 163-182. https://doi.org/10.1016/S1060-3743(96)90023-5
Suleiman, M. F. (2000). *The process and product of writing: Implications for elementary school teachers.* The California Association for Bilingual Education Conference.
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
Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997
Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*(3), 305-319. https://doi.org/10.1177/026553229301000306
Zohaib, M., Memon, S., & Syed, H. (2021). Academic Writing Development in English: An Action Research Project. *Elementary Education Online, 20*(5), 3193-3204. 10.17051/ilkonline.2021.05.347

Acknowledgements

Being a two-month visiting scholar at York University, Toronto, was beneficial for my academic and professional development. I owe a tremendous debt of gratitude and humbleness to Pr. Khalid Barkaoui from York University in Canada introduced me to the Multi-Facet Rasch Model (MFRM) and the statistical program FACETS. He, thankfully, walked me gradually through the basics of this program to be able to analyse its output tables and answered my copious questions about the operation of FACETS during Rasch analysis. Moreover, I am immensely grateful to all library staff and teachers in the Institute of Education in London and Sussex University in Brighton for helping me to have free access to various articles, theses, and books in my field of expertise.

Funding
Not applicable.

Ethics Declarations

Competing Interests
No, there are no conflicting interests.

Rights and Permissions