



www.EUROKD.com

Language Testing in Focus: An International Journal



Language Testing
in Focus
An International Journal
LTiF



ISSN : 2717-9087

2024 (10)

The inter-rater and intra-rater reliability in scoring writing tests in UTAS

Mahdi Rouhiathar*, Katelyn Howard

English Language Unit, Preparatory Studies Centre, University of Technology and Applied Sciences, Salalah, Sultanate of Oman

ABSTRACT

Keywords

Inter-rater Reliability, Intra-rater Reliability, UTAS, Writing Test

Received

09 May 2024

Received in revised form

21 August 2024

Accepted

05 September 2024

Correspondence concerning this article should be addressed to:

atharmehdi@gmail.com

Rater reliability, the consistency of marking across different raters and times, is one important component of reliability regarding the quality of test scores. It is essential regarding performance assessment, such as writing, when the fairness of assessment results can come into question due to the subjectivity when scoring. The present study, part of a larger-scale funded research project, aimed to study this overlooked area in the Omani context, i.e., the reliability in scoring the writing section of the final exams in the University of Technology and Applied Sciences (UTAS). More specifically, the study investigated the estimates of inter-rater and intra-rater reliability among 10 writing markers assessing 286 and 156 students' writing scripts belonging to four levels of proficiency at three different levels of analysis: the whole writing tests, tasks 1 and tasks 2, and the constituent criteria of both tasks. The results indicated a rather high value of inter-rater reliability and a moderate one for intra-rater reliability in general. However, when interpreted regarding raters' personal and background information, some low estimates shed light on the importance of factors influencing scoring consistency across different assessors and times.

Introduction

Writing clearly and effectively is essential in personal, academic, and professional areas. Due to its importance, it is a critical skill used in performance assessment in first and second language studies. Written tasks, such as paragraphs or essays, are a requirement in nearly all

standard language tests. These may include proficiency and achievement assessments at the college and university level. In Oman's Universities of Technology and Applied Sciences (UTAS), written essays are required at all levels (i.e., 1-4). However, the rating of writing assessments is a complex endeavour. In both first- and second-language contexts, the subjectivity of rating has attracted a lot of attention.

Raters must understand and use four important qualities to have the most accurate results for test takers. These qualities are validity, reliability, impact, and practicality (Bachman, 1990). Reliability, while an essential part of validity and a potential source for proof of construct validity, has been a difficult term to define. However, Jones (2012) suggested that reliability includes four essential ideas: consistency, error, generalizability, and dependability. In his opinion, a reliable test should consistently yield the same or similar results when used repeatedly, should ideally be free from errors with the variations in scores being primarily due to the measured ability and not other factors, should be able to reproduce the testing scenario in a manner that offers evidence, and should be reliable in terms of consistency in classification.

The reliable rating of a set of written scripts is the essential standard at which scoring should be set. The assessment involves ensuring that various raters and ratings at different time points would result in consistent ratings for a group of students. The group of students would be rated on their writing ability regardless of the variability regarding student or task variation, individual circumstances, raters' characteristics or experience, and test issues. Along with this idea, rater reliability is a crucial concept to get consistent scores or marks. As of now, human ratings are used in rating written performances. However, humans are also a huge source of error based on their subjective ratings.

There are two ways to look more closely at this problem of rater reliability: inter-rater and intra-rater reliability. The first one is concerned with the agreement on the rating of a performance by two different raters (Shohamy, 1983). When an assessment is given and rated, the ratings should be consistent regardless of which rater rates the performance. If this is true, then students will receive a reliable rating and will not be concerned with whom will mark their performance (Fulcher, 2003). Intra-rater reliability is a measure of the level of agreement when a single rater conducts numerous evaluations, as a component of rating consistency. When assessing a particular language performance, whether it be spoken or written, an examiner uses specific criteria, as outlined by Bachman (1990). When the rater consistently applies the same criteria to evaluate the language skills of various test takers, it leads to a reliable collection of ratings.

Along with rater reliability, another source of concern is the rubrics of the assessments themselves (Nakatsuhara, Khabbazbashi, & Inoue, 2022). Rubrics are usually a set of levels with descriptive criteria at each level that are usually in a hierarchy. The learner is then measured and placed at each level depending on the descriptive criteria. At each level in the rubric, there are descriptors. These descriptors are what the assessment developers affirm to be rating or assessing (Fulcher, 1996; Davies et al., 1999). The majority of rubrics are created for

raters' use to give an accurate rating for learner performance based on the level descriptors. These are called 'examiner-oriented' scales.

The rating scale applied in UTAS, according to the current assessment policies, tends to be an analytic, multiple-trait scale with an ability focus. It is adopted and adapted from the one employed in IELTS. According to UTAS assessment manuals (2022), the construct of written performance encompasses a series of constituent micro constructs (i.e., task achievement, coherence & cohesion, task response, organization, lexical resources, grammatical range & accuracy). These are verbally described across different bands (i.e., 0-5 for levels 1-3 and 0-10 for level 4).

Ideally, there must be no problems with subjectivity with the clear-cut boundaries between the bands. However, it has proved not to be the case and, accordingly, there have been some inconsistencies in scoring. The evaluation appears to involve listing particular performance characteristics, but the evaluator typically leans towards making an overall, comprehensive assessment of performance, resulting in minimal counting or tracking of features or mistakes in most cases. Separate ratings for each trait or descriptor are the underlying assumption when using the multiple-trait rating scale. Nevertheless, in actuality, the evaluator usually and inadvertently tends to form a singular opinion regarding the performance of a specific construct, such as 'communicative ability'. The ability to focus is another trait of a rubric that will have descriptors that contain the skills needed for a test taker to complete the assessment successfully. These are usually also able to describe a successful task completion in a real-world context. However, this could lead a rater to focus too much on real-world context and what an examinee can do rather than following the rubric descriptors. Finally, a rater may subconsciously evaluate performances based on the preceding and following performances instead of looking at the rating scales.

Concerning the above-mentioned issues in scoring the examinees' writing performances, the researchers attempted to measure the inter-rater and intra-rater reliability of the examinees' writing performance at three different levels of analysis and to interpret the results in terms of the raters' personal characteristics and background information.

Literature Review

Varying features of this topic have been written about in the literature and a brief review follows. The reliability of written assessments and their ratings has been a difficult task for decades. With human raters, there will always be variation in what raters focus on and prefer when rating. Many factors are causing this variation (Kayapınar, 2014; Stuart, 2023). The problems of inter-rater/intra-rater reliability, the significant aspects affecting them, and their measurement have also been well-researched (Bachman, 1990; McNamara, 2000; Brown, 2004; Luoma, 2004; Weir, 2005; Sak, 2008; Brown, 2011; Yen, 2016; Sureeyatanapas et al, 2024).

Language assessment in general and oral/written assessment in particular have been underrepresented in the studies done in Oman. Though partly related to the area of interest in

the current study, it is worthy of note that Al Hajri (2014), in her analysis of English language assessment in the Colleges of Applied Sciences in Oman, addressed inconsistency in carrying out assessment criteria. The writer talked about how important it is for the College of Applied Sciences (CAS) to have reliable assessment documents and the involvement of accreditation and quality assurance agencies in encouraging academic institutions to use reliable measures of achievement, as well as explaining the methods used to guarantee uniformity in applying these measures. She mentioned that the English Department at CAS released three policy papers in 2009, 2010, and 2011, according to the standards for assessment uniformity and regulation. Each document indicated that they did not meet the criteria for giving accurate evaluations of student work. Many of these problems were due to the diverse levels and great workload of those in coordination positions (Al Hajri, 2014).

The scarcity of studies concerning the topic in the Omani context in general and UTAS, in particular, necessitates a systematic exploration of reliability in writing tests. With these issues in mind, one can surprisingly find no prior research attempts addressing the critical issue of subjectivity in marking UTAS students' written performance in Foundation courses despite its central role in test fairness and washback effect. Although several measures have been taken to minimize the sources of errors concerning raters' variations, it is essential to explore inter- and intra-rater reliability in a more systematic and quantifiable way. Accordingly, the researchers addressed the issue by measuring rater consistency in a hierarchy of levels, beginning with the whole writing test, continuing with its two constituent sections (i.e., task 1 and task 2), and ending with the micro constructs of the criteria (i.e., task achievement, task response, organization, coherence and cohesion, vocabulary/lexical resources, grammar/grammatical range, and accuracy) (UTAS, 2022, pp 20-21).

With this innovative and unprecedented methodology and following the objectives delineated in the previous part, the following tentative main questions with the follow-up ones were proposed and addressed:

RQ1: Is the measure of inter-rater reliability at the UTAS writing test satisfactory?

RQ2: Is the level of inter-rater reliability at the UTAS writing test satisfactory in task 1 and task 2?

RQ3: Is the level of inter-rater reliability at the UTAS writing test satisfactory in the micro constructs of the criteria?

RQ4: Is the measure of intra-rater reliability at the UTAS writing test satisfactory?

RQ5: Is the level of intra-rater reliability at UTAS writing test satisfactory in task 1 and task 2?

RQ6: Is the level of intra-rater reliability at the UTAS writing test satisfactory in the micro constructs of the criteria?

Method

Concerning the research questions, one can see that the design of the present research is descriptive or non-experimental and correlational. This study applied a quantitative method. In the first stage, the researchers used the scores that the raters gave to different essays to analyze and investigate inter-rater consistency at three different hierarchical levels. In the second stage,

the raters' evaluation of the same essays at two different times was employed to measure intra-rater reliability at three different levels of hierarchy. Accordingly, different parts of the methodology, namely the subjects (statistical population), the materials (rating scales), the data collection procedure, and the data analysis are presented and discussed.

Participants

In line with the aims of this research project to estimate the level of inter- and intra-rater reliability on written English proficiency tests, 10 university teachers who were assigned to mark the writing scripts voluntarily participated in marking the final exam writing scripts independently. The raters' demographic information is presented in the following table:

Table 1
Markers' General Experience with Assessment

Rater	Gender	NS/NNS	Age	Tertiary	UTAS	Level 1	Level 2	Level 3	Level 4	Assessment Coordinator	IELTS and TOEFL	Prior Assessment Training
1	F	NS	30	4.5	2	0	1	1	1	0	No	Yes
2	M	NS	44	12	7	1	1	2	3	0	No	No
3	M	NNS	34	10	1	0	1	1	0	0	No	Yes
4	M	NS	50	15	12	2	3	2	3	2	Yes	Yes
5	M	NNS	43	10	1	2	2	1	1	0	Yes	Yes
6	M	NNS	28	2	1	2	1	1	1	0	Yes	Yes
7	M	NS	33	3.5	2.5	3.5	3.5	3.5	3.5	0	Yes	Yes
8	M	NNS	53	31	22	9	9	5	12	0	Yes	Yes
9	F	NNS	50	25	2	1	1	1	1	0	Yes	Yes
10	F	NS	62	7	7	4	1	1	5	0	Yes	Yes
		Avg:	43	12	5.8	2.45	2.35	1.85	3.05	0.2	70% Yes 30 % No	90% Yes 10% No

The distribution of the examinees across the four levels of proficiency can also be seen in the following tables:

Table 2
Examinees' Levels 1 – 4 (Inter-rater Reliability)

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Total
Level 1	13	-	-	-	-	-	13
Level 2	28	19	29	24	23	28	151
Level 3	24	20	25	12	-	-	81
Level 4	19	22	-	-	-	-	41
						Total	286

Table 3*Examinees' Levels 1 – 4 (Intra-rater Reliability)*

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Total
Level 1	6	6	-	-	-	-	12
Level 2	12	12	12	12	12	12	72
Level 3	12	12	12	12	-	-	48
Level 4	12	12	-	-	-	-	24
						Total	156

Instruments

The raters who participated in the study used four assessment booklets as reference while marking the scripts. They include all detailed test specifications: writing learning outcomes, exam elements (format, time, the number of tasks, time, expected answer, mark assignment, and penalty), the corresponding learning outcomes, writing marking criteria (i.e., task achievement, task response, coherence and cohesion, lexical resource, grammatical range and accuracy), detailed marking writing rubrics for the two tasks, marking procedure for the midterm exam and final exam (UTAS, 2022, pp. 20-22).

Procedures

The writing component, lasting for 60 to 70 minutes depending on the level, consists of two academic-focused tasks; Writing Task 1 and Writing Task 2. It is suggested that about 20 minutes be spent on Task 1. The first task is more prescriptive as it gives students a finite number of items that must be included in their responses. Each level has different tasks that could include writing an email, an incident report, a chart, or a table description. On the other hand, the second task usually requires the candidate to write an organized paragraph or essay (depending on the level). Task 2 should take about 40 minutes. The candidates are assessed on their ability to follow a paragraph or essay structure on a standard academic topic.

In task 1, test takers are assessed based on the following criteria: 1. Task fulfilment 2. Organization/coherence and cohesion 3. Vocabulary/lexical resources 4. Grammar/grammatical range and accuracy. In task 2, test takers are assessed based on the following criteria: 1. Task response 2. Organization/coherence and cohesion 3. Vocabulary/lexical resources 4. Grammar/grammatical range and accuracy. It is worthy of note that the terms used to refer to the criteria differ across different levels. For instance, the terms *coherence and cohesion* are used for the level 4 rubric while the term *organization* is employed in those of levels 1, 2, and 3 (UTAS, 2022).

Concerning the estimation of inter-rater reliability, emails were sent to the markers after the exam, giving them instructions about what was expected from them. More specifically, the researchers thanked the participants for their support with the research data collection procedure. They were told to use the blind marks of Marker 1 and Marker 2 sheets while marking the writing papers for the foundation exams. They were also reminded that this would not be the moderated marks that they agreed upon. In other words, they had mark sheets given to them as they began marking. It was emphasized that they should mark as they normally would and give the researchers the blind mark sheets before they moderate together. Finally,

they were ensured that all records would be kept confidential, anonymous, and used for research purposes only. The mark sheets including the raters' names and their marks across the two tasks and criteria were duly collected after administering the exam by taking the maximal measures to observe the blind marking policy.

The procedure for the second stage, intra-rater reliability, lasted for six months. Six papers from each marker's rated papers were chosen to be marked again by the same rater after five months. It is worthy of note that the papers were selected from low-, average- and high-achievers and were copied with personal details and marks eradicated. The emails were sent to the markers, and they were asked to mark the selected papers again.

Data Analysis

The correlation coefficient is a statistical tool employed to assess the connection between two variables and can be used to calculate rater reliability (Luoma, 2004). To do this, data was gathered which included scores from a test, scores from two assessors, and scores from one assessor at two different time points. The sets of scores came from the same test and test-takers, but two different raters and two different times. Out of the possible ways to calculate the correlation between the two sets of scores, Pearson r was employed to calculate both inter- and intra-rater reliability after the tabulation of the resultant scores in Excel sheets. The obtained scores were coded in the three levels of the whole test, task 1 and task 2, and the marking criteria (e.g., task achievement) mentioned above.

Results

The calculations of correlation values were done for the whole tests, tasks 1 and 2 across different levels and groups, the results of which can be observed in Table 4.

Table 4

The Correlation Estimates of the Whole Tests, Tasks 1 and Tasks 2 across Different Levels and Groups

Level	Group	Total	Task 1	Task 2
1	1	0.9	0.74	0.94
2	1	0.31	0.34	0.42
2	2	0.87	0.59	0.88
2	3	0.79	0.9	0.81
2	4	0.81	0.81	0.79
2	5	0.97	0.97	0.98
2	6	0.97	0.95	0.98
3	1	0.81	0.74	0.89
3	2	0.76	0.64	0.87
3	3	0.86	0.78	0.79
3	4	0.45	0.72	0.22
4	1	0.94	0.89	0.98
4	2	0.76	0.72	0.84
	Mean	0.78	0.75	0.80

The average overall correlation agreements between the two raters were 0.78, 0.75, and 0.80 for the whole tests, tasks 1 and tasks 2, respectively. Based on Table 4 above, the findings show that the rating, in general, was reliable and there were good agreements between the raters as they awarded similar marks to the students. Concerning the third level of inter-rater reliability estimation, the resultant data can be observed in the following table.

Table 5
Correlations between Different Criteria of Rubrics

Level	Group	Task 1				Task 2			
		Task Achievement	Organization	Grammar	Vocabulary	Task Response	Organization	Grammar	Vocabulary
1	1	0.89	0.77	0.71	0.61	0.79	0.84	0.91	0.83
2	1	0.25	0.38	0.30	0.24	0.48	0.73	0.34	0.44
2	2	0.79	0.93	0.70	0.90	0.85	0.82	0.82	0.89
2	3	0.56	0.60	0.66	0.65	0.61	0.76	0.49	0.54
2	4	0.80	0.73	0.72	0.77	0.79	0.75	0.79	0.80
2	5	0.86	0.96	0.86	0.97	0.80	0.77	0.91	0.66
2	6	0.88	0.86	0.87	0.83	0.71	0.89	0.86	0.93
3	1	0.69	0.44	0.67	0.73	0.82	0.79	0.72	0.81
3	2	0.40	0.53	0.27	0.38	0.89	0.60	0.58	0.66
3	3	0.54	0.55	0.60	0.54	0.75	0.87	0.62	0.67
3	4	0.70	0.21	0.63	0.75	0.27	0.05	0.05	0.26
4	1	0.74	0.62	0.76	0.76	0.87	0.93	0.92	0.92
4	2	0.70	0.74	0.71	0.74	0.87	0.85	0.70	0.65
	<i>Mean</i>	0.68	0.64	0.65	0.68	0.73	0.74	0.67	0.70

As can be seen, most of the values show that the agreement between raters is high and statistically significant. However, the estimates are comparatively lower than those of the first two levels (i.e., the whole test and tasks 1 and 2). Concerning the measures of intra-rater reliability, the researchers employed three levels of analysis like the ones used in the inter-rater reliability. The results of the first two levels, the whole test, tasks 1 and 2 can be observed in Table 6.

According to the table, the estimates indicate significant correlations between the two ratings of the same marker. However, there are lower measures in comparison to those of inter-rater reliability. The correlation estimates of the third level of analysis, different criteria of the rubrics, are presented in Table 7.

Table 6

The Correlation Estimates of the Whole Tests, Tasks 1 and Tasks 2 across Different Levels and Markers

Level	Marker	Total	Task 1	Task 2
1	1	0.76	0.67	0.79
1	2	0.79	0.79	0.74
2	1	0.74	0.68	0.78
2	2	0.76	0.75	0.73
2	3	0.58	0.63	0.13
2	4	0.44	0.50	0.67
2	5	0.56	0.48	0.48
2	6	0.43	0.58	0.74
3	1	0.72	0.70	0.74
3	2	0.72	0.71	0.67
3	3	0.61	0.67	0.70
3	4	0.68	0.65	0.59
4	1	0.61	0.56	0.56
4	2	0.68	0.49	0.62
	Mean	0.64	0.63	0.63

Table 7

Correlations between Different Criteria of Rubrics

Level	Marker	Task 1				Task 2			
		Task Achievement	Organization	Grammar	Vocabulary	Task Response	Organization	Grammar	Vocabulary
1	1	0.58	0.61	0.54	0.53	0.88	0.88	0.76	0.78
1	2	0.78	0.78	0.78	0.78	0.68	0.75	0.76	0.77
2	1	0.64	0.67	0.68	0.33	0.46	0.47	0.31	0.41
2	2	0.75	0.79	0.74	0.66	0.74	0.77	0.67	0.72
2	3	0.45	0.37	0.56	0.47	0.55	0.45	0.45	0.33
2	4	0.44	0.50	0.40	0.61	0.56	0.55	0.64	0.60
2	5	0.41	0.55	0.44	0.56	0.45	0.56	0.45	0.66
2	6	0.54	0.60	0.56	0.67	0.56	0.45	0.54	0.76
3	1	0.57	0.67	0.47	0.76	0.62	0.56	0.56	0.77
3	2	0.58	0.60	0.59	0.70	0.60	0.61	0.59	0.78
3	3	0.61	0.56	0.56	0.80	0.45	0.46	0.56	0.76
3	4	0.59	0.62	0.60	0.82	0.45	0.67	0.76	0.80
4	1	0.55	0.60	0.58	0.76	0.54	0.60	0.59	0.75
4	2	0.60	0.56	0.61	0.77	0.55	0.59	0.70	0.68
	Mean	0.57	0.61	0.58	0.66	0.58	0.60	0.60	0.68

Despite the significance of agreement between the two ratings in most cases, the estimates at the third level are lower in comparison to those of the first two levels. The lowest agreement can be seen in task achievement due to the vagueness of the criterion in the rubrics.

Interestingly, the second lowest agreement was related to grammar due to the differences between an NS's and an NNS's conceptualization of the criterion. In contrast, a relatively higher agreement existed in the assessment of vocabulary.

Discussion

According to the results concerning inter-rater reliability, the average estimates demonstrate high-reliability values. However, there can be observed very few cases where the agreement between the raters was low. This can be due to several factors to be discussed later. Out of the three averages shown in the results, the highest one belonged to task 2 while that of task 1 had the last ranking. This may indicate that the scoring became more reliable as the raters proceeded from task 1 to task 2. In other words, the inferences made by the raters about the examinees' performance become progressively more accurate and reliable.

The disagreement among raters, though very low, could be due to inaccuracy and inconsistencies in marking as raters have varying experience in teaching and marking writing at the tertiary level in general and UTAS in particular. Regarding the raters' personal information and linguistic, academic, and social background, it can be claimed that the sample adequately represents the population as it includes raters of different ages (i.e., 28-62), linguistic backgrounds (i.e., NS/NNS), genders, teaching and marking experience at tertiary level inside and outside of UTAS, IELTS/TOEFL marking experience and assessment training experience. Despite the differences, the association between the two raters would be considered statistically significant by normal standards. In other words, inter-rater consistency estimates are higher than what the researchers expected, indicating that the inferences made about the examinees' writing performance were probably accurate and reliable.

The results of the third level of analysis indicate that measuring inter-rater reliability at different levels can give us more accurate ideas about rater consistency. Furthermore, the obtained means demonstrate that the similar micro constructs (e.g., grammar) steadily rose from task 1 to those of task 2. This can be indicative of the fact that the inferences made by raters become progressively more accurate as the raters proceed from task 1 to task 2.

At this level, the raters' personal characteristics and background information contributed more to the differences between their assessment of the examinees' performance. In this regard, more experienced raters (e.g., *level 2, group 5*) show the highest agreement. The pairing of NS raters resulted in more agreement about *task achievement* and *organization* (e.g., *level 2, group 6*) while pairing of an NS and an NNS as raters led to wide differences in terms of *grammar* (e.g., *level 3, group 2*). According to the results, there seems to be more agreement when both members of the pairs are male or female (e.g., *level 1, group 1*).

Concerning the results of intra-rater reliability, there were more cases where the agreement between the two ratings were lower than .5 and, accordingly, the raters' rating was not reliable (e.g., *level 2*). Interestingly, the highest agreement belonged to level 1 raters, while the lowest ones were observed among level 2 teachers. Like the findings concerning inter-rater reliability,

varying experience in teaching and marking writing at the tertiary level in general and UTAS in particular was the main cause of disagreement. In contrast to the estimates of inter-rater reliability, not a high level of variation among markers was observed while moving from task 1 to task 2.

When it comes to the third level of the analysis in the second stage, the lowest agreement can be seen in task achievement due to the vagueness of the criterion in the rubrics. Interestingly, the second lowest agreement was related to grammar due to the differences between an NS's and an NNS's conceptualization of the criterion. In contrast, a relatively higher agreement existed in the assessment of vocabulary.

The findings of the study demonstrate that more attention should be paid to the pairing of raters by considering the differences between them. This needs a more considerate arrangement of raters according to influential factors such as experience, age, gender, ethnicity, etc. The other possible factors that can contribute to the differences between the raters can be the evaluator's severity or leniency, social and linguistic background, prior training in assessment, NR/CR orientation, teaching and rating experience, fatigue, adherence to rubrics' criteria, work pressure for deadline and responsibility for accuracy (Yen, 2016). More specifically, the raters' reading styles, scoring methods, and the differences between the reading order are among other elements that can affect raters' evaluations of examinee performances.

It is worthy of note that some raters seem to be equally aware of all the criteria while others seem to prioritize more on one or two of the criteria. This can be because the rating scales need to be clarified in terms of vague terms by extra training, multiple marking, and consensus moderation to reduce inconsistencies. This can be organized by appointing an experienced colleague as a chief examiner (CE).

The assessment workshops and training sessions are currently perceived to be less effective than expected in reducing the sources of error threatening the reliability of the tests. This is reflected in the informal talks among English teachers as they regard the sessions as a waste of time. The number of workshops held for the lecturers should be increased. More importantly, the quality of these sessions should also be improved to reduce the inconsistencies in scoring the learners' written performances.


Conclusion

Regarding the findings of the study, it can be concluded that the rater consistency in UTAS writing exams is relatively high, indicating that the agreement between the raters is noticeable. However, there seem to be more observable differences between raters when the other levels of consistency are also examined. This means that the differences between raters in terms of personal characteristics and linguistic/academic background can account for the differences between the inferences made and the resultant low estimates of inter-rater reliability at different levels of analysis.

English as the only official foreign language within the Sultanate of Oman enjoys a unique status as it is taught in most government schools from the first grade, is the dominant medium of instruction at the tertiary level, and is in great demand by the job market. Given this, there should be efforts to standardize the rating of students' written performance and reduce the inconsistency in the administration and marking of such exams. This endeavour aimed to study the inter-rater reliability measures in UTAS quantitatively to shed light on the current situation. This analysis helps the decision-makers at all levels to seek possible problems and solutions. Future studies can focus on replicating the current study by choosing samples from the other branches of UTAS to see whether the results are confirmed or rejected. In addition, the other form of reliability, intra-rater reliability, can be considered and examined regarding raters' backgrounds. Chiefly, the factors contributing to the rater consistency can be investigated to help increase the reliability of inferences made by the assessors.

ORCID

 <https://orcid.org/0000-0001-9951-3903>

 <https://orcid.org/0009-0000-6865-309X>

Acknowledgments

Firstly, we should acknowledge the Scientific Research Department at the University of Technology and Applied Sciences (CAS) Salalah for funding this research. We would also like to acknowledge the leadership and colleagues within the English Department for their support in this research endeavour.

Funding

Not applicable.

Ethics Declarations

Competing Interests

No, there are no conflicting interests.

Rights and Permissions

Open Access

This article is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/), which grants permission to use, share, adapt, distribute, and reproduce in any medium or format provided that proper credit is given to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if any changes were made.

References

- Al-Hajri, F. (2014). English assessment in the Colleges of Applied Sciences in Oman: Thematic document analysis. *English Language Teaching*, 7(3), 19-37. 10.5539/elt.v7n3p19
- Al-Mahrooqi, R., & Denman, C. (2018). English Language Proficiency and Communicative Competence in Oman: Implications for Employability and Sustainable Development. In *English Language Education* (pp. 181-193). (English Language Education; Vol. 15). Springer Science and Business Media B.V. https://doi.org/10.1007/978-981-13-0265-7_11
- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford University Press.
- Brown, A. (2012). Interlocutor and rater training. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of Language Testing* (pp. 413-525). Routledge.
- Brown, H. (2004). *Language assessment: Principle and classroom practice*. Pearson.
- Brown, H. D. & Abeywickrama, P. (2018). *Language assessment: Principles and classroom Practices* (3rd ed.). Pearson Education.
- Brown, J. D. (2011). *Testing in language programs: A comprehensive guide to English language assessment*. McGraw Hill.Education.

- CAS (2011). *Assessment policies*. Ministry of Higher Education.
- Davidson, F. (2012). Test specifications and criterion-referenced assessment. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of Language Testing* (pp. 427-439). Routledge.
- Davies, A., Brown, A., Edler, C., Hill, K., Lumley, T. & McNamara, T. (1999). *Dictionary of language testing: Studies in language testing 7*. Cambridge University Press.
- Feldt, L. S. and Brennan, R. L. (1989). Reliability. In R. L. Linn (ed.) *Educational Measurement* (pp.105–146). American Council on Education/Macmillan.
- Fulcher, G. (1996a). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, 13(2), 208-238. <https://doi.org/10.1177/026553229601300205>
- Fulcher, G. (2003). *Testing second language speaking*. Longman/Pearson Education.
- Fulcher G, Harding L (2022). *Routledge handbook of language testing*. Routledge.
- Isaziga, V. A. A. & Ruiz, D. M. (2019). *Estimating inter-rater reliability on an oral English proficiency test from a bilingual education program*. Universidad de Tecnológica de Pereira. <https://repositorio.utp.edu.co/server/api/core/bitstreams/24845692-6f12-47d8-800b-a24c49027347/content>
- Jones, N. (2012). Reliability and dependability. In G. Fulcher & F. Davidson (Eds.) *The Routledge Handbook of Language Testing* (pp. 350-362). Routledge.
- Jönsson, A., & Thornberg, P. (2014). Samsyn eller samstämmighet? En diskussion om sambedömning som redskap för likvärdig bedömning i skolan. *Pedagogisk forskning i Sverige*, 19(4-5), 386–402.
- Kang, O., & Rubin, Don. (2012). Intra-rater reliability of oral proficiency ratings. *The International Journal of Educational and Psychological Assessment*, 12(1), 43-61. https://www.researchgate.net/publication/329504535_Intra-rater_Reliability_of_Oral_Proficiency_Ratings
- Karavas, E., & Delieza, X. (2009). On-site observation of KPG oral examiners: Implications for oral examiner training and evaluation. *Apples-Journal of Applied Language Studies*, 3(1), 51–77. <https://apples.journal.fi/article/view/97803>
- Kayapınar, U. (2014). Measuring essay assessment: Intra-rater and inter-rater reliability. *Eurasian Journal of Educational Research*, 57, 113-136. [www. https://dergipark.org.tr/en/pub/ejer/issue/5164/70223](http://www.dergipark.org.tr/en/pub/ejer/issue/5164/70223)
- Klenowski, V., & Adie, L. E. (2009). Moderation as judgment practice: Reconciling system level accountability and local level practice. *Curriculum Perspectives*, 29(1), 10–28. https://www.researchgate.net/publication/27483149_Moderation_as_judgment_practice_Reconciling_system_level_accountability_and_local_level_practice
- Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing*, 19, 246 - 276. <https://doi.org/10.1191/0265532202lt230oa>
- Luoma, Sari. (2004). *Assessing speaking*. Cambridge University Press.
- McNamara, T. F. (2000). *Language testing*. Oxford University Press.
- Nakatsuhara F, Khabbazbashi N, & Inoue, C. (2022). Assessing speaking. In G. Fulcher & L. Harding (eds.), *Routledge Handbook of Language Testing* (pp. 209-222). Routledge. <http://hdl.handle.net/10547/625318>
- Rashid, S., & Mahmood, N. (2020). High stake testing: Factors affecting inter-rater reliability in scoring of secondary school examination. *Bulletin of Education and Research*, 42(2), 163-179. <https://files.eric.ed.gov/fulltext/EJ1280726.pdf>
- Sadler, D. R. (2013). Assuring academic achievement standards: from moderation to calibration. *Assessment in Education: Principles, Policy & Practice*, 20(1), 5–19. <http://dx.doi.org/10.1080/0969594X.2012.714742>. doi:10.1080/0969594X.2012.714742
- Sak, G. (2008). *An investigation of the validity and reliability of the speaking exam at a Turkish university* [M.A. - Master of Arts]. Middle East Technical University. <http://etd.lib.metu.edu.tr/upload/3/12610080/index.pdf>
- Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th language testing research colloquium* (Vol. 9, pp. 129–152). Orlando, Florida. Cambridge: Cambridge University Press.
- Shohamy, E. (1983) Rater reliability of the oral interview speaking test. *Foreign Language Annals*, 16(3), 219-222. <https://doi.org/10.1111/j.1944-9720.1983.tb01456.x>
- Stanley, G., MacCann, R., Gardner, J., Reynolds, L., & Wild, I. (2009). *Review of Teacher Assessment: Evidence of What Works Best and Issues for Development*. University of Oxford Centre for Educational Assessment. <http://hdl.handle.net/1893/32425>
- Stuart, N. J., & Barnett, A. L. (2023). The writing quality scale (WQS): A new tool to identify writing difficulties in students. *British Journal of Special Education*, 1–10. <https://doi.org/10.1111/1467-8578.12464>
- Sundqvist, P., Sandlund, E., Skar, G. B., & Tengberg, M. (2020). Effects of rater training on the assessment of L2 English oral proficiency. *Nordic Journal of Modern Language Methodology*, 8(1), 3-29. <https://doi.org/10.46364/njmlm.v8i1.605>

- Sureeyatanapas, P., Sureeyatanapas, P., Panitanarak, U. (2024). The analysis of marking reliability through the approach of gauge repeatability and reproducibility (GR&R) study: a case of English-speaking test. *Language Testing in Asia* 14, 1. <https://doi.org/10.1186/s40468-023-00271-z>
- UTAS (2022). *General foundation program - English language: Level 1-4 testing specifications*. Ministry of Higher Education.
- Veerappan, V. & Sulaiman, T. (2012). A review on IELTS writing test, its test results and inter-rater reliability. *Theory and Practice in Language Studies*, 2(1), pp. 138-143. <https://doi.org/10.4304/tpls.2.1.138-143>
- Wang, P. (2009). The inter-rater reliability in scoring composition. *English Language Teaching*, 2(3), 39-43. <http://doi.org/10.5539/elt.v2n3p39>
- Weigle, S.C. (2002). *Assessing writing*. Cambridge University Press.
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.
- Yen, N. T. Q. (2016). Rater consistency in rating L2 learners' writing task. *VNU Journal of Foreign Studies*, 32(2), 75-84. <https://jfs.ulis.vnu.edu.vn/index.php/fs/article/view/1544/1506>