Language Related Research
E-ISSN: 2383-0816
https://lrr.modares.ac.ir
https://doi.org/10.29252/LRR.12.5.24
http://dorl.net/dor/20.1001.1.23223081.1400.12.5.24.8

Tarbiat Modares University

Vol. 12, No. 5
pp. 623-652
November & December 2021

Received: 12 June 2021
Received in revised form: 26 July 2021
Accepted: 4 September 2021

On the Accuracy of English Language Teachers' Writing Assessment

*Masoud Azizi**

Abstract

In case not enough caution is exercised in the assessment of second or foreign language learners' writing performance, one cannot trust the accuracy of decisions made accordingly. As experts or trained raters are often not available or it is not cost-effective to employ them in most educational contexts, writing assessment is often carried out by language instructors, who may not enjoy an adequate competence in teaching and assessing L2 writing. This makes the investigation of the accuracy of ratings done by language teachers a must. In so doing, 30 language teachers in three groups, each with a different background in teaching English and L2 writing, were selected, and their ratings of 30 IELTS samples were compared against those of expert raters using One-Way ANOVA tests. A statistically significant difference was found among the raters for the total writing score as well as the four components, with the L2 writing teachers demonstrating the closest performance to that of the expert rater and with language teachers with no or very little background in teaching L2 writing demonstrating the lowest accuracy. Moreover, the only significant correlations were found between the ratings done by the writing teachers and those of the expert rater, indicating that only they could interpret the scoring criteria not significantly different from the expert rater. The results demonstrate that language teachers are not generally suitable writing raters as they are affected by their own teaching background and understanding of the rating criteria.

Keywords: L2 writing, assessment, rater training, teacher raters, rating accuracy

* Corresponding author: Assistant Professor, Department of Foreign Languages, Amirkabir University of Technology, Tehran, Iran; *Email: mazizi@aut.ac.ir*, ORCID ID: https://orcid.org/0000-0001-9054-1131

1. Introduction

Performance assessment in general and writing assessment, in particular, are necessarily intertwined with subjective judgments of the test takers' performance, which can raise questions regarding the validity and reliability of such an assessment (Weigle, 2002). Due to the multifaceted nature of the writing skill and the rubrics often employed for assessment, tackling this issue is of great significance.

Rating students' writing samples as the most prevalent method of evaluation in assessing writing has always been accompanied by the question of reliability, concerns for the construct-irrelevant factors, and test fairness. Both researching and evaluating students' writing performance often involve ratings of their samples using either a holistic or analytic writing scale or rubric (Weigle, 2002). However, no matter how well such raters have been trained, they are likely to be affected by a number of construct irrelevant variables. When it comes to novice raters or especially language teachers rather than raters, the situation gets even worse as teachers' assessment literacy often lags behind their teaching competence (Crusan et al., 2016).

In most educational settings, this is the language teachers rather than expert raters who are responsible for carrying out the required assessment in the case of productive skills. Teachers who have not even received any training on how to teach writing, for instance, have to assess their students' writing samples during or at the end of the program. Language institutions often tend to do so as the use of expert raters is often not cost-effective. However, when language teachers who have not received any kind of training in teaching and assessing writing are employed for such an assessment, one should exercise a great deal of caution when basing her decisions on the obtained results. The situation is even worse when we learn that the majority of language teacher preparation programs ignore the need for incorporating a component of teaching and assessing writing (Hodges et al., 2019).

Language teachers are often expected to carry out the assessment at least for the program they teach themselves. Even in low-stakes tests, accuracy in the scores assigned to students' writing samples is of great importance as inaccurate judgments of learners' performance at the end of a program may question the consequential validity of the test because it can put the learners under pressure both mentally and financially. It implies the learners' admission of failure, and it imposes more time, money, and endeavor on the part of the learners. However, not every learner is motivated enough to be able to deal with such a situation (Dornyei, 2001) especially when she feels her failure was the result of inaccurate judgments. This is quite opposite of what learners expect their teachers to be (Estaji & Zhaleh, 2021). There might be students who quit learning a foreign or second language (Dornyei & Ushioda, 2011).

Imagine a language institute in which three or more language teachers are teaching classes of the same level. At the end of the program, each teacher is responsible for the assessment of his or her own students. For reading and listening skills, there is no problem as similar questions with objective scoring can be employed, but for speaking and writing which are productive skills, the judgments are by nature quite subjective (Bachman & Palmer, 2010). Even if the same scoring rubric is given to the teachers to rate students' writing samples, for example, still there is no guarantee that they interpret the rubric similarly and come up with comparable ratings. This may question test fairness.

Being a good teacher does not necessarily mean being a good rater. One may know how to teach a language, but she does not necessarily know how to assess it. Like all raters, teachers may be affected by their own characteristics which are often considered construct-irrelevant variables. In addition, achieving a high level of reliability, consistency, and accuracy may be very challenging as they are less likely to have received the required training. Studies indicate that the majority of teachers consider themselves inadequately prepared for the assessment of their students' performance (Crusan et al., 2016; Mertler, 2009). They feel ill-equipped in

performance assessment, which makes them feel uncomfortable and unprepared (Zhu, 2004).

In spite of the significance of writing instruction and assessment in education, very few teacher preparation programs incorporate a component of writing instruction and assessment (Hodges et al., 2019; Martin & Dismuke, 2018). According to Myers et al. (2016), only 25% of preservice teachers take a course in writing instruction.

Research shows that writing teachers often lack the required assessment literacy (Crusan et al., 2016). Hirvela and Belcher (2007) assert that the field of L2 writing has ignored the need for preparation of L2 writing teachers in assessment. Instead, it has centered all its attention on students' learning of the L2 writing skill. Weigle (2007) too, acknowledges the need for training L2 writing teachers in assessment.

There have been a number of studies examining the accuracy of teachers' assessments (e.g., Attali, 2015; Ecks, 2008; Engelhard & MyFord, 2003; Jolle, 2014; Wind & Engelhard, 2013), with some specifically focusing on the effect of teachers' background and experience on their judgments (e.g., Brown et al., 2004; Lim, 2011; Wiseman, 2012), but to the best of the author's knowledge, none has compared general language teachers with L2 writing teachers' assessments against those of expert raters. Moreover, the inclusion of language teachers with different background in familiarity with L2 writing conventions can be regarded as one of the merits which differentiate this study from others. In so doing, the present study investigated the accuracy of scores assigned by language teachers in general, with variability in their familiarity with L2 writing conventions, and writing instructors in particular against those of trained raters. In so doing, the following research questions were stated:

1. How accurate are language teachers in rating students' writing performance?
2. To what extent are teacher raters dealing with the same construct as expert

raters when assessing students' writing samples?

2. Review of the Literature

Teaching has always been intertwined with evaluation. Teaching writing is no exception. Writing teachers have to do some form of evaluation. Even when they are providing their students with corrective feedback, they are in fact evaluating their students' performance. When it comes to assessment, i.e. evaluation in the form of quantification (Bachman & Palmer, 2010), the picture gets even more complicated. Assigning a score to a writing sample is of great significance as it involves subjective judgments and at the same time needs to truly represent the student's writing ability. In other words, it must represent the construct it is supposed to measure. Moreover, it needs to be consistent, i.e., it does not change over time and the same or at least very similar score is awarded to the same sample if it is rated by more than one rater.

It is very important that we make sure the scores we award students' performance with are so reliable that they can back up the decisions we make accordingly. It is crucial that we make sure we are in fact assessing the construct we intend to; otherwise, the inferences we make accordingly could be questionable. In addition to the concerns about the construct validity of our assessment, consistency in such an assessment is imperative especially in the context of high-stakes tests (Bachman & Palmer, 2010).

As Attali (2015, p. 1) indicates, "a primary goal is to ensure that raters think similarly enough about what characteristics of student responses determine their quality to achieve reasonable consistency of scores across ratings." However, achieving such a goal is not an easy task. Raters' performance has been found to reveal great variability due to the way raters interpret the scoring criteria (Attali, 2015; Ecks, 2008; Engelhard & MyFord, 2003).

Performance assessment scores are mediated through human …. As a result,

interpretation and use of rater-mediated descriptions of a student's performance depends on the rater's ability to interpret and apply the rating scale as intended. In order to trust ratings as useful descriptions of student performance, it is necessary to systematically examine the quality of ratings (Wind & Engelhard, 2013, p. 279).

When it comes to the writing assessment, two crucial factors stand out; the scoring construct and the rater quality (Jolle, 2014; Suto, 2012). Regarding the former, the rating method, i.e., holistic vs. analytic (Barkaoui, 2011) and the type of prompt given (Weigle, 1999) have been observed to affect raters' performance in the scoring process. In the case of the latter, rater background (Lim, 2011; Wiseman, 2012), rater style (Ecks, 2008), training (Brown et al., 2004), scoring criteria (Clauser, 2000), and the raters' values and expectations (Baker, 2010) are only some of the variables affecting the rater in the scoring process.

Research shows that raters do not pay adequate attention to common frameworks or scoring rubrics; instead, they tend to rely on their own tacit knowledge while scoring (Jolle, 2014). As Barkaoui (2007, p. 105) states: "rater factors - such as personality, cultural, linguistic, and education background, and rating experience - influence rater decision-making behavior, interpretations and expectations concerning task requirements and scoring criteria, reaction to ESL/EFL essays, severity (inter-rater reliability) and self-consistency (intra-rater reliability)."

The scores raters assign to students' writing samples are under the influence of many factors including their background and experience in rating, teaching writing, and prior language learning experience (Barkaoui, 2010; Weigle, 2002). "The textual features of an essay, the wording of the rating scale, and all the impressions readers bring with them – as well as the potential interaction of these elements" – may have an impact on the raters' perception of the writing sample they are reading and affect their scores consequently (Goodwin, 2016, p. 2).

Among all the variables affecting raters, experience may be the most widely

explored factor as it is possible to be tackled through training (Weigle, 2016). Raters' familiarity with L2 writing and the conventions of students' L1 writing has been found to increase raters' sympathy toward such writings (Shaw & Weir, 2007). In addition, raters' academic discipline is another variable having an impact on raters' scoring. Research indicates that to disciplinary faculty, content is more important than language. English Composition teachers are more concerned with the notion of audience in comparison with other specialists, while grammar and cohesion are more important to ESL/EFL teachers than other specialists (Shaw & Weir, 2007).

The quality of the samples previously rated can also affect a rater's judgment of the sample being reviewed. Raters have been reported to score an average quality writing sample lower if they had rated a high-quality sample right before it (Goodwin, 2016). The two proceeding rated samples of highly different quality can cause a strong rater bias in the case of the sample being rated in comparison with the time when only one contrasting quality sample exists before it (Spear, 1997).

Wolfe et al. (2016), synthesizing various components of the rating process contributing to rater inaccuracy and inconsistency, propose four general categories of variables affecting raters while scoring. First, the design of the assessment may have a clear impact on the quality of the ratings. This is about the decisions made by the assessment designers and may include issues such as "the purpose of the assessment, the administration medium, and the focus of the scoring criteria" (p. 2). Second, the content of the response given by the test takers may affect the scores raters assign. By content, though, Wolfe et al. (2016, p. 2) mean "the visual appearance of the response (e.g., handwriting quality, font choices, and page layout), textual features (e.g., length, word choice, mechanics), and content included in the response (e.g., author clues, ideas)." Third is rater characteristics including their "experience (e.g., educational, demographic, and professional), stable rater cognitive and affective traits (e.g., temperament, cognitive style), and temporary rater states (e.g., mood)." Finally, the fourth category is the rating

context which encompasses "the medium and process through which responses are distributed to raters, rater training procedures, rater monitoring and feedback practices, and temporal and physical features of the rating environment" (p.3).

According to Wolfe and McVay (2012), there are two lines of research in the case of rater-mediated assessment, with one focusing on the impact rater characteristics such as experience, expertise, and training might have on the scoring process and the other one focusing on the statistical procedures to check and control the quality of ratings. These statistical procedures and indices are often classified in terms of three main categories: "(1) rater agreement, (2) rater errors and systematic biases, and (3) rater accuracy" (Wind & Engelhard, 2013, p. 280).

Regarding the rater agreement, researchers check the extent to which raters assign matching scores to the same sample of performance. Such indices include categorical agreement and measures of association among raters. In the case of the second category, i.e. rater errors and systematic biases, one is concerned with the specific patterns in the rating process that may lead to the assignment of scores that are different from those representing a students' performance. Finally, rater accuracy is described as the "match between operational ratings and those established as 'true' or 'known' ratings by individual or committees of expert raters" (Wind & Engelhard, 2013, p. 280). In other words, the extent to which the ratings of a rater match those of an expert rater determines the accuracy of his ratings. Therefore, rater accuracy is often defined as a comparison between the scores given by raters and those regarded as the standard, often assigned by trained and expert raters (Wang et al., 2017). As such, a high level of agreement with a low level of errors and biases, and together with a high level of accuracy are believed to reflect a high-quality rating (Wind & Engelhard, 2013).

According to Brown et al. (2004), in order to determine score accuracy and consistency in writing assessment, consensus estimates and consistency estimates are often employed. While consensus measures are used to determine the extent to

which raters assign the same scores, consistency measures represent the degree to which the pattern of high and low scores is similar between and among the raters.

Consensus measures are often used when the markers are trained to rate based on a specific scale rubric. This approach includes indices of percent exact agreement, which is the percentage of the scores which are exactly identical between the raters, and percent adjacent agreement, i.e. the percentage of scores that fall within plus and minus one score category or band from each other (Brown et al., 2004). Obviously, the adjacent agreement measure yields higher indices than the exact agreement index, and with scales of few categories, usually up to four bands, achieving higher indices is much easier (Stemler, 2004). Achieving an exact agreement score of 70% or higher is considered to indicate a reliable scoring (Brown et al., 2004). Achieving exact agreement even when the scale is quite short is very difficult. Therefore, adjacent agreement indices are considered to be more robust especially when the rubric is adequately long (Brown et al., 2004). A review of the studies in writing by Brown et al. (2004) indicated that an exact agreement rate of 40% to 60% and an adjacent agreement rate of 80% to 100% are often commonly reported in studies. Moreover, consistency coefficients of between .70 to .80 are often observed for standardized performance tests in writing when all the students respond to the same prompt.

The existence of a noticeable pattern in the distribution of scores assigned by raters can be checked using Pearson coefficient for pairs of raters and Cronboch's alpha for multiple raters as measures of consistency coefficients. A high coefficient indicates that the raters gave high and low scores in a similar pattern. On the other hand, it is possible that one obtains a high correlation between two raters while their means are quite different from each other as high for two raters does not necessarily mean the same scores. Similarly, when there is very little variance among the scores of the two raters, i.e., they assign very similar scores with high agreement indices, a very low coefficient may be obtained, which could result in drawing the false conclusion of having a low reliability (Brown et al.,

2004). In addition, this measure can also indicate the extent to which two measures are dealing with the same construct. The coefficient of determination, i.e., the correlation coefficient squared, represents the common variance between two measures or the extent to which two measures are tapping the same construct. That is why in the present study, this measure has been employed to answer the second research question.

The present study was an attempt to examine the extent to which language teachers' assessment of learners' writings matches that of the expert raters. It was an attempt to check the effect of experience or expertise in teaching second language writing on teacher raters' rating accuracy.

3. Methodology

3.1. Participants

Based on experience and expertise in second language writing, three groups of teachers were identified for the purpose of the present study. For each category, ten teachers with similar characteristics matching the criteria for each group, were selected using snowball sampling. They all had a minimum experience of 15 years in English language teaching. All three groups were university instructors who also taught general English courses. They were all PhD holders of TEFL with an age range of 35 to 43. Twenty-one were male and 9 were female. Finally, three in each group were randomly selected to take part in the study. The reason why only three teachers in each group were selected for participation was that each group's ratings had to constitute one set of scores to be included in data analysis for the purpose of the comparison with other groups, and the maximum number of raters in the evaluation of inter-rater reliability indices is often three. In case more ratings were obtained from teachers in each group, it was not possible to come up with one set of scores accordingly. In this study, for the ratings done by teachers in each group, the scores assigned by the teacher with the highest index of intra-

rater reliability were selected for data analysis. In case the difference between his and those of the teacher with the next highest intra-rater reliability was more than one band score, the average between the two closest ratings was selected to be included in data analysis for that specific writing sample.

The first group of teachers (hereafter called R1), besides teaching a range of academic courses at universities and writing journal articles, had a rich background in teaching writing but had received no official training in writing assessment. The second group of teachers (R2) mainly taught general English at universities but had a rich background in discourse analysis and writing journal articles. They had received no official training in teaching and assessing writing either. Finally, the third group of teachers (R3) was mainly language teachers with very little experience in dealing with academic writing. They were mainly involved in teaching general English at language institutes with periodical teaching of some general English courses at universities. What differentiated these groups were their experience and expertise in teaching writing (R1) and familiarity with L2 academic writing conventions (R1 & R2). Group three actually lacked the two criteria and were only experienced in teaching English as a foreign language.

It is worth mentioning that all the teachers, as PhD holders in TEFL, had, for sure, had courses in language assessment during their MA and PhD programs, but those did not include any practical and specific components on assessing L2 writing.

3.2. Materials

In order to check the accuracy of ratings, 30 samples of IELTS writing task 2 already rated by IELTS official raters were used to be assigned to the participants to rate. In addition, IELTS Writing Task 2 scoring rubric was employed to help participant raters to score the given samples. All the samples were of different prompts.

3.3. Data Collection Procedure

The present study was an attempt to check the accuracy of ratings done by language teachers and the extent to which it is possible to trust the ratings done by untrained teacher raters. In order to do so, three groups of teachers were selected: one group consisted of university instructors with experience in teaching second language writing and academic writing but with no official training in assessing L2 writing as self-reported by the participants and checked by the researcher to see if that fits their resume and field of work. The second group encompassed the university instructors with no formal experience in teaching academic L2 writing, but quite experienced in writing academic papers and familiar with the conventions of academic writing. The third group included language teachers who had been mainly concerned with teaching English as a foreign language in language institutes and universities. For this group of teachers, L2 writing had been restricted to the periodic exercises available in the usual textbooks available on the market. Ten teachers were identified in each group and from among them, three were randomly selected for participation in the study. They had not been trained in rating writing based on the IELTS scoring rubric, and their familiarity with the IELTS test was limited to the periodic teaching of the IELTS courses in the past.

Thirty actual IELTS writing samples for task 2 in the writing section of the test were obtained. These samples were written in actual IELTS test sessions and were all rated by official IELTS raters. The scores available were those of the four components checked in the scoring rubric namely, Task Achievement, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy. The total score was the mean of these four components.

The participant raters were given the IELTS writing task 2 scoring rubric and the 30 samples to rate accordingly. No training was given as it was attempted to check if language teachers could be good raters by nature. They were asked to provide the researcher with one score for each of the four scoring components. The scores could range between 0 and 9. No half score was possible. The half-

band scores reported in the IELTS test are due to averaging the scores given for the four components in the rubric. The mean scores, which could be in the form of half-band scores, were later calculated by the researcher.

The scores of the three groups of raters were compared with those of the expert raters to see if their assessments were significantly different. In addition, According to Brown et al. (2004), for determining score accuracy and consistency in writing assessment, consensus estimates and consistency estimates should be checked. Consensus measures are used to examine the extent to which raters assign the same or similar scores while consistency measures indicate the degree to which the pattern of high and low scores is similar between and among the raters. On the other hand, consensus estimates are often used when the raters are trained to rate based on a specific scale rubric. This approach includes indices of percent exact agreement, and percent adjacent agreement. Moreover, the existence of a pattern in the distribution of scores can be studied using the Pearson coefficient for pairs of raters as measures of consistency coefficients. Since the participants in the present study were not trained in assessing writing, only the consistency measures, as well as their general performance, were examined.

3.4. Data Analysis

In order to compare the performance of the four groups of raters and answering the first research question, a one-way ANOVA was used for each of the four components in the scoring rubric as well as the total score for the writing task. In addition, to check the consistency measures and answer the second research question, the Pearson-Product Moment coefficient was employed.

4. Results

The results of the One-Way ANOVAs run among the four groups of ratings showed a significant difference for the overall scores as well as the four components. However, the post hoc analysis showed that the picture was not quite similar in each case.

In the case of the overall scores obtained from the four ratings, a significant difference was found among the raters, f (3, 106) = 5.56, p = .00. The multiple comparisons using the Tukey test showed that the only significant differences were between Rater 1 and 2 (p = .02), and Rater 1 and 3 (p = .00). Table 1 presents the descriptive statistics for this measure. As evident in Table 1, the mean for the performance of Rater 1is the closest mean to that of the Expert rater while the means for Rater 2 and 3 are the closest to each other.

Interestingly enough, the analysis of means for homogenous subsets, provided as part of the Tukey result output, categorized the Expert Rater and Rater 1 in one subset, while Rater 2 and 3 were put in another subset together with the Expert Rater, indicating that the performance of Rater 1 was quite different from that of the other two participant raters.

Table 1
Descriptive Statistics for Raters' Overall Score

	N	Min.	Max.	Mean	SD	Std. Error
Expert Rater	30	5	8	6.42	.821	.150
Rater 1	30	6	9	6.68	.942	.172
Rater 2	30	5	7	6.05	.724	.162
Rater 3	30	5	7	5.97	.392	.072
Total	120	5	9	6.30	.796	.076

Regarding the IELTS writing rubric components, a significant difference was observed among the raters in the case of the Task Achievement component, f (3, 116) = 5.10, p = .00. Regarding the means, Table 2 indicates that Rater 1 has the closest performance to that of the expert rater.

Table 2
Descriptive Statistics for Raters' Scores on Task Achievement Component

	N	Min.	Max.	Mean	SD	Std. Error
Expert Rater	30	3	9	6.90	1.647	.301
Rater 1	30	5	9	6.77	1.073	.196
Rater 2	30	4	7	6.05	1.050	.235
Rater 3	30	5	7	5.90	.712	.130
Total	120	3	9	6.44	1.245	.119

As the results of the post hoc analysis showed, the first rater's performance on this task was very similar to that of the expert rater (p = .97) while the other two were quite different with a trend being observed between the expert rater and Rater 2 (p = .07) and a significant difference between the expert rater and Rater 3 (p = .01). In the case of the teacher raters, a significant difference was observed between Rater 1 and 3 (p = .03). Rater 2 and 3's performance was found to be very similar (p = .97).

In the output specifying homogeneous group subsets, Rater 2 and 3 were put in one group, Rater 1 and 2 in another group, and Rater 1 and the Expert Rater in a different group, indicating that Rater 1 could significantly outperform other teacher raters and perform more similarly to the expert rater.

In the case of the second component, i.e. Coherence and Cohesion, a significant difference was observed among the raters, f (3, 116) = 4.52, p = .00. As Table 3 indicates, the performances of Rater 1 and 2 were more similar to that of the expert

rater in comparison with that of Rater 3. The results of the Tukey HSD test, however, showed that none of the teacher raters significantly differed from the expert rater, and the only significant difference lay between Rater 1 and Rater 3 (p = .00).

Table 3
Descriptive Statistics for Raters' Scores on Coherence and Cohesion Component

	N	Min.	Max.	Mean	SD	Std. Error
Expert Rater	30	5	9	6.27	1.015	.185
Rater 1	30	5	9	6.57	.971	.177
Rater 2	30	4	7	6.05	.826	.185
Rater 3	30	5	7	5.77	.568	.104
Total	120	4	9	6.17	.907	.087

However, the picture was completely different in the case of the Lexical Resource component. The results of the One-Way ANOVA run among the raters' given scores indicated a significant difference, f (3, 116) = 6.53, p < .005, with the difference between Rater 1 and the Expert Rater being almost none (p = .99) and the difference between the Expert Rater and Rater 2 (p = .02) and Rater 3 (p = .01) being significant. Also, the differences between Rater 1 and 2 (p = .03) and Rater 1 and 3 (p = .01) were found statistically significant. However, virtually no difference was found between Rater 2 and 3 (p = 1.00).

Table 4
Descriptive Statistics for Raters' Scores on Lexical Resource Component

	N	Min.	Max.	Mean	SD	Std. Error
Expert Rater	30	5	8	6.50	.861	.157
Rater 1	30	5	8	6.47	.973	.178
Rater 2	30	5	7	5.80	.768	.172

	N	Min.	Max.	Mean	SD	Std. Error
Rater 3	30	5	7	5.80	.551	.101
Total	120	5	8	6.17	.866	.083

Finally, regarding the Grammatical Range and Accuracy component, a significant difference was observed when running a One-Way ANOVA test, f (3, 116) = 3.43, p = .02. However, the post hoc tests showed no significant difference among the raters, with only a trend being observed between the Expert Rater and Rater 2 (p = .07). While the difference between the raters was not significant based on the Tukey test, the observed p could indicate that they were not very much similar. The observed p for the difference between the Expert Rater and Rater 1 was .99, while it was .13 between the Expert Rater and Rater 3. The p value for the difference between Rater 1 and 2 was .10, between Rater 1 and 3 was .17, and between Rater 2 and 3 was .97, indicating that Rater 2 and 3 had a very similar performance while that of Rater 1 was less similar. Table 5 presents the related descriptive statistics. This could be due to the conservativeness of the Tukey test in comparison with other tests of similar function.

Table 5

Descriptive Statistics for Raters' Scores on Grammatical Range and Accuracy Component

	N	Min.	Max.	Mean	SD	Std. Error
Expert Rater	30	5	9	6.37	.964	.176
Rater 1	30	5	9	6.33	1.093	.200
Rater 2	30	4	7	5.75	.786	.176
Rater 3	30	5	7	5.87	.507	.093
Total	120	4	9	6.11	.902	.086

While measures that check the existence of differences in performance may serve well in many cases, they may not be able to present a vivid picture of the situation we are dealing with as they work with means rather than individual scores. Another set of measures often employed to indicate if raters are actually measuring the same construct is association measures. In order to answer the second research question and to check the correlation between the ratings of the raters, the Pearson Product Moment correlation coefficient was employed as the scorings were in rank order scale. Table 6 presents the strength of association between the scores given by the expert rater and those of the teacher raters. The ones found statistically significant are marked by an asterisk.

Table 6
Strength of Association Between the Expert Rater and the Teacher Raters

	Rater 1	Rater 2	Rater 3
Overall Score	.492*	.059	.070
Task Response	.449*	.272	.129
Coherence & Cohesion	.062	-.242	.196
Lexical Resource	.383*	.169	.059
Grammatical Range & Accuracy	.514*	.237	.270

* Correlation is significant at the 0.05 level (2-tailed).

As evident in the above table, the only rater who could have a similar performance to that of the expert rater was rater 1. However, that seemed not to be true in the case of the Coherence and Cohesion component. In addition, even for Rater 1, the coefficients found were not large enough. In fact, to have a more vivid picture regarding the extent to which the raters were dealing with the same

construct, one need to square the correlation coefficients to come up with the common variance between the two measures. In that case, in the best case scenario, there was only about 25% shared variance between the performance of the Expert rater and Rater 1, which indicates that though writing teachers enjoy a better position in comparison with other language teachers in assessing second or foreign language writings, they still may not be the right choice for such an assessment if they do not receive any training. However, due to their better performance in assessment in comparison with other groups of teachers, they could be good candidates for training in writing assessment.

5. Discussion

A statistically significant difference was observed among the ratings done by the Expert Rater and the three teacher rater groups in the total scores. While all raters' total scores were not significantly different from those of the expert rater in this regard, the performance of the writing teachers (Rater 1) was found significantly different from those of the Rater 2 and 3. However, what is of more significance is the evaluation of the four components as the total score is the mean score for those components. The analyses done showed that there was great variability in teachers' performance in assessment in the case of different components. The fact that a significant difference was observed between the ratings of Rater 1 and those of the other two teacher groups can imply that experience and expertise in teaching L2 writing can cause variations in teachers' assessment. This affirms Barkaoui (2010) and Weigle's (2002) observations in this regard.

In the case of the components in general, and Task Achievement in specific, the differences among the performance of teacher raters manifested themselves in a more vivid manner. While Rater 1 performed similarly to the Expert Rater, the other two rater groups showed a significant difference. Interestingly enough, as evident in the tables presenting the descriptive statistics for each component,

while the maximum score in the scores assigned by the Expert Rater and Rater 1 was 9, it was only 7 in the scorings done by Rater 2 and 3, indicating less experienced teacher raters' expectations of always finding some flaws in the text written by students and that there could be no perfection in at least a simple component of the test. This could also be due to the hegemony of IELTS test and teachers' assumption that scoring 9 in such a test or at least some of its components is often out of reach, more specifically when they encounter some errors in the candidates' writings as they often expect perfection for a full score in any test. This is in line with Baker's (2010) findings that the raters' expectations and values can affect their assessment in the scoring process. This could also justify why Raters 2 and 3 were more conservative in assigning scores while Rater 1 was not. That could be why Rater 1 was categorized with the Expert Rater in one group while the other two teacher raters were put in a different category based on the means for groups in homogenous subsets.

In the case of the Coherence and Cohesion component, the performance of all raters was similar to that of the Expert Rater though that of Rater1 was closer. This, however, contradicts the results of the consistency measures as none of the correlations between the scores given by teacher raters and those of the Expert Rater was found significant, which indicates that though the general performance examined in terms of the mean scores was not very much different, the teacher raters were actually dealing with a different construct in comparison with the Expert Rater. This could have different reasons but one could be the difficulty in the interpretation of the IELTS scoring rubric for this component. First of all, other elements such as grammatical accuracy are more straightforward concepts that allow a simpler conversion to more practical and tangible indications in the texts written by students. However, cohesion and coherence do not lend themselves well to such a conversion and their interpretation could vary from teacher to teacher. More importantly, when examining the scoring rubric on this component, one can observe that these concepts are not defined in the rubric but

they are often simply named especially for the higher band scores. The use of phrases and sentences such as "uses cohesion in such a way that it attracts no attention" (Band Score 9) and "manages all aspects of cohesion well" (Band Score 8) can indicate the subjectivity of the rater's interpretation of this component. This is in line with other studies (e.g., Attali, 2015; Ecks, 2008; Englehard & MyFord, 2003) reporting great variability in raters' performance as the result of differences in the way they interpret the scoring criteria. It also confirms Goodwin's (2016) assertion that even the wording of the rating scale can affect the raters' perception of a writing sample and impact the scores they assign as a result.

Regarding the Lexical Resource component, while Rater 1's performance was almost exactly the same as that of the Expert Rater, the other two groups of teacher raters performed almost exactly the same as each other but significantly different from those of the Expert Rater and Rater 1. This might be again due to the raters' different interpretations of the rubric or even differences in the raters' own range of vocabulary, which could further affect their interpretation of the rubric especially when it reads "uses a wide range of vocabulary with very natural and sophisticated control of lexical features" (Band Score 9), "uses a wide range of vocabulary fluently and flexibly to convey precise meanings" (Band Score 8), "skillfully uses uncommon lexical items" (Band Score 8), "uses a sufficient range of vocabulary to allow some flexibility and precision" (Band Score 7), or "uses less common lexical items with some awareness of style and collocation" (Band Score 7). As Wind and Engelhard (2013) assert, the interpretation of the scoring criteria by raters to a great extent depends on the rater's ability to interpret and apply the intended scale. "Performance assessment scores are mediated through human raters who exist within complex ecological contexts" (p. 279).

Finally, in the case of the Grammatical Range and Accuracy, while the performances of Rater 1 and Expert Rater were very similar in comparison with that of the other teacher raters, no significant difference was observed between the Expert Rater and the teacher raters. That could be attributed to the

straightforwardness of this component and the ease of interpretation of the scoring rubric in this case. It seems safe to assume that teachers will know what an error-free piece of writing is and can distinguish between simple and complex sentence structures used by the student writers. This confirms Clauser's (2000) observation that the scoring criteria itself is one important factor affecting raters' performance.

In addition, the closeness in performance between Rater 1 and the Expert Rater but not those of other teacher raters was confirmed by the examination of the measures of consistency. The only significant correlations were found between those of the Expert Rater and Rater 1 but not the other two teacher raters, though the observed correlations were not large enough. Since only Rater 1 had a rich background in teaching academic writing, this could confirm Shaw and Weir (2007) and Weigle's (2016) observation that not only experience makes a difference in raters' performance, but their familiarity with L2 writing conventions can significantly affect their ratings. This also affirms Wolfe et al.'s (2016) categorization of factors affecting raters as the results of the present study fall into their third category, i.e. rater characteristics including their educational, demographical and professional experience.

The differences in performance among the teacher raters indicate that their expertise and background in teaching L2 writing, and their familiarity with the principles of academic writing, as a result, can significantly affect their writing assessment performance. In other words, the more familiar teachers are with L2 writing conventions, the more logical it is to expect them to have a less subjective assessment. This is in line with the results of Lim (2011) and Wiseman (2012) indicating that teacher experience and their teaching background do affect the scores they assign to student writing samples. In addition, the similarity in performance between Rater 2 and 3, who did not enjoy a profound background in teaching writing, indicates that a rich background in teaching academic writing may be able to assist teachers to enjoy a better assessment skill though even such an advantage cannot guarantee a very accurate scoring on the part of the teachers.

Moreover, the differences between the performance of teacher raters and that of the expert rater can imply that all teachers including L2 writing instructors need training in the assessment if obtaining an accurate set of scores is the objective. That is why Brown et al. (2014) emphasize on the importance and the effect of training on raters' performance. In addition, it seems that such training may be more successful if received by teachers already enjoying a rich background in L2 writing.

Finally, the fact that very low correlations were obtained between ratings done by the expert rater and those of the teacher raters indicates that the way teachers interpret the same scoring rubric is not necessarily similar. Each might actually be dealing with a different construct as they judge students' performance under the influence of their own experience and understanding of what a good piece of writing should look like (Jolle, 2014). That is why Attali (2015) emphasizes on the importance of ensuring that raters think similarly enough about what student writing characteristics should be taken into account in assessment so that a high level of consistency and accuracy is obtained.

6. Conclusion

The sensitivity of assessment in most educational settings entails that appropriate measures be taken in order to ensure the accuracy of the ratings done. Language instruction is of no exception. Today, internationally accredited language tests such as IELTS and TOEFL have an important role to play in our professional development no matter what educational background we come from. In addition, language instruction at language institutions or in a more academic setting has always been accompanied by some type of assessment as it is often the summative assessments at the end of a program that determines if one has successfully completed that program.

In case not enough caution is exercised in the implementation of such

assessments, one cannot trust the accuracy of the decisions made on the basis of such ratings. This is of greater importance when it comes to high-stakes tests. Assessing writing is of equal, if not more, importance as performance assessment has always been associated with subjective judgments on the part of the raters (Bachman & Palmer, 2010). Therefore, what is evident is the need to ensure that the assessment of students' performance is as accurate as possible. However, often expert raters are not available or it is not practical or cost-effective to make use of expert raters in most educational contexts. As a result, performance assessments are often carried out by language instructors, who were found not to be suitable for such an assignment if not trained.

The results of the present study indicate that language teachers are not suitable writing raters as they are affected by their own experience in teaching and their understanding of the rating criteria. General language teachers, even those to some extent familiar with L2 academic writing conventions, often lack the necessary competence in teaching and assessing writing (Crusan et al., 2016). These teachers may act like novice teachers in decision-making in the field of assessment (Khatib & Saeedian, 2021). The only group of teachers who could, to a large extent, reach a level of performance in assessment to that of the expert raters, though not in all cases, was L2 writing teachers. General language teachers with limited experience in teaching L2 writing and even those familiar with L2 academic writing conventions, but not teaching L2 writing professionally, were unable to demonstrate an acceptable level of performance in comparison with not only expert raters but also writing teachers. This designates the important role of teaching L2 writing as an intervening variable in assessing this skill. Even being able to write well in English or being familiar with its conventions through discourse analysis, as in the case of the third group of teacher raters, cannot guarantee accuracy in writing assessment.

The present study implies that the teacher training programs need to incorporate not only courses on how to teach writing but also components on how

to assess it. Teachers need instruction on how to avoid variables negatively affecting their judgments of students' performance as well as on how to interpret a scoring rubric similarly. Policymakers and institutional authorities need to invest in their teachers' assessment literacy if they wish their decisions to be based on accurate assessments.

Moreover, these findings indicate that for any test whose results may make a difference for individuals, i.e., where consequential validity matters, we need to make use of either expert raters or at least trained teachers with a long experience in teaching and assessing L2 writing. The use of teachers, especially when two or more teachers are responsible for the assessment of students who are supposed to be judged based on the same criteria, can jeopardize the accuracy of the ratings and question the quality of decisions made on their basis.

Still another implication this study may have is for those in rater training. While all individuals may have the potential for becoming a good writing rater with training, those already involved and experienced in teaching academic writing are better candidates and should be of priority.

The present study, as all studies do, faces a number of limitations. The variable under investigation in the present study was teachers' experience and expertise in teaching writing, while raters may be affected by a large number of factors whose investigation was beyond the scope of the present study. In addition, the scoring rubric used in the present study was that of the IELTS writing component which is a nine-band descriptor. When the number of bands increases, it would be much more difficult to come up with very accurate scorings especially if the raters have not been trained for the use of that rubric. The use of other rubrics with fewer bands may result in different findings, which can be investigated in future studies. Moreover, for the second research question, due to the scope of the study, only a quantitative analysis was done. A more thorough and qualitative analysis of the reasons behind teachers' variability in the interpretation of the rubric needs to be carried out to shed further light on the topic.

References

Attali, Y. (2015). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, *1*, 1–7. https://doi.org/10.1177/0265532215582283

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.

Baker, B. A. (2010). Playing with the stakes: A consideration of an aspect of the social context of a gate keeping writing assessment. *Assessing Writing, 15*, 133–153. https://doi.org/10.1016/j.asw.2010.06.002

Barkaoui, K. (2007). Participants, texts and processes in ESL/EFL essay tests: A narrative review of the literature. *Canadian Modern Language Review, 64*(1), 99–134.

Barkaoui, K. (2010). Do ESL essay raters' evaluation criteria change with experience? A mixed-method crosses-sectional study. *TESOL Quarterly, 44*(1), 31–57. https://**doi**.org/10.5054/tq.2010.214047

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing, 28*, 51–75. https://doi.org/10.1177/0265532210376379

Brown, G., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing, 9*, 105–121. https://doi.org/10.1016/j.asw.2004.07.001

Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological, Measurement, 24*(4), 310–324. https://doi.org/10.1177/01466210022031778

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices.

Assessing Writing, 28, 43–56. https://doi.org/10.1016/j.asw.2016.03.001

Dornyei, Z. (2001). *Motivational strategies in the language classroom*. Cambridge University Press.

Dorneyi, Z. & Ushioda, E. (2011). *Teaching and researching motivation* (2nd ed.). Harlow.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*, 155–185. https://doi.org/10.1177/0265532207086780

Engelhard, G., Jr., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition Program with a many-faceted Rasch model* (College Board Research Report No. 2003–1). College Entrance Examination Board.

Estaji, M. & Zhaleh, K. (2021). Exploring Iranian teachers' perceptions of classroom justice and its dimensions in EFL instructional contexts. *Language Related Research, 12*(3). 277–314. https://doi.org/10.29252/LRR.12.3.10

Goodwin, S. (2016). A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes. *Assessing Writing, 30*, 21–31. https://doi.org/10.1016/j.asw.2016.07.004

Hirvela, A., & Belcher, D. (2007). Writing scholars as teacher educators: Exploring writing teacher education. *Journal of Second Language Writing, 16* (3), 125–128. https://doi.org/10.1016/j.jslw.2007.08.001

Hodges, T. S., Wright, K. L., Wind, S. A., Matthews, S. D., Zimmer. W. K., McTigue, E. (2019). Developing and examining validity evidence for the Writing Rubric to Inform Teacher Educators (WRITE). *Assessing Writing, 40*, 1–13. https://doi.org/ 10.1016/j.asw.2019.03.001

Jolle, L. (2014). Pair assessment of pupil writing: A dialogic approach for studying

the development of rater competence. *Assessing Writing, 20*, 37–52. https://doi.org/ 10.1016/j.asw.2014.01.002

Khatib, B., & Saeedian, A. (2021). Identifying and informing novice Iranian English language teachers' classroom decision making and pedagogical reasoning regarding managerial mode. *Language Related Research, 12*(3). 121–149. https://doi.org/10.29252/LRR.12.3.5

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*, 543–560. https://doi.org/10.1177/0265532211406422

Martin, S. D., & Dismuke, S. (2018). Investigating differences in teacher practices through a complexity theory lens: The influence of teacher education. *Journal of Teacher Education, 69*(1), 22–39. https://doi.org/10.1177/0022487117702573

Mertler, C. (2009). Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. *Improving Schools, 12*(1), 101–113. https://doi.org/10.1177/1365480209105575

Myers, J., Scales, R. Q., Grisham, D. L., Wolsey, T. D., Dismuke, S., Smetana, L., et al. (2016). What about writing? A national exploratory study of writing instruction in teacher preparation programs. *Literacy Research and Instruction, 55*(4), 309–330. https://doi.org/10.1080/19388071.2016.1198442

Shaw, S. & Weir, C. (2007). *Examining writing: Research and practice in assessing second language writing* (Vol. 26). Cambridge University Press.

Spear, M. (1997). The influence of contrast effects upon teachers' marks. *Educational Research, 39*(2), 229–233. https://doi.org/10.1080/0013188970390209

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment Research &*

Evaluation, 9(4). 1-11. https://doi.org/10.7275/96jp-xz07

Suto, I. (2012). A critical review of some qualitative research methods used to explore rater cognition. *Educational Measurement: Issues and Practice, 31*, 21–30. https://doi.org/10.1111/j.1745-3992.2012.00240.x

Wang, J., Engelhard Jr G., Raczynski, K., Song, T., & Wolfe, E.W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing, 33*, 36–47. https://doi.org/10.1111/j.1745-3984.1996.tb00479.x

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

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

Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing, 16*(3), 194–209. https://dx.doi.org/10.1016/j.jslw.2007.07.004

Weigle, S. C. (2016). Second language writing assessment. In R. M. Manchón & P. K. Matsuda (Eds.). *Handbook of second and foreign language writing* (pp. 473–494). De Gruyter Mouton.

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

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing, 17*, 150–173. https://doi.org/10.1016/j.asw.2011.12.001

Wolfe, E.W., & McVay, A. (2012). Application of latent trait models to identifying substantially interesting raters. *Educational Measurement: Issues and Practices, 31*(3), 31–37. https://doi.org/10.1111/j.1745-3992.2012.00241.x

Wolfe, E.W., Song, T., & Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing, 27*, 1–10. http://dx.doi.org/10.1016/j.asw.2015.06.002

Zhu, W. (2004). Faculty views on the importance of writing, the nature of academic writing, and teaching and responding to writing in the disciplines. *Journal of Second Language Writing, 13*(1), 29–48. https://doi.org/10.1016/j.jslw.2004.04.004

About the Author

Masoud Azizi is an assistant professor of Applied Linguistics and the Head of the Foreign Languages Department at Amirkabir University of Technology. His main areas of interest are teaching and assessing writing.