THE JOURNAL OF ASIA TEFL
Vol. 16, No. 2, Summer 2019, 735-742
http://dx.doi.org/10.18823/asiatefl.2019.16.2.22.735

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511

Promoting Fairness in EFL Writing Assessment: Are There Any Effects of the Writers' Awareness of the Rating Criteria?

Nasim Ghanbari
Persian Gulf University, Bushehr, Iran

Introduction

In traditional language testing, test-takers are tested on knowledge that has already been defined and shaped by testers who are in authority. As an isolated activity, tests are developed with little consideration for the stakeholders involved (Lynch, 1997; Shohamy, 2001). This is rooted in the fact that use of the test in the context and for a particular group of test-takers is not the concern here. Rather, the development of psychometrically sound tests which embody the intentions of their developers is strongly pursued. Apparently, in this unequal power relationship, test takers have no voice and they have to merely comply with the decisions made by testers. A critical view of language testing, however, questions the passive role of test-takers who have been long viewed as black boxes in the testing process (Lynch, 2001; Shohamy, 2001).

The solution posed by the critical language testing (CLT) movement to improve such unethical practice is to involve test-takers as the most important stakeholders in the testing enterprise (Rea-Dickins, 1997). Involving test-takers at different stages of testing can improve the validity of the assessment, democratize the testing process and consequently it can have a positive impact on the learning process (Yu, 2007).

Inspired by the CLT movement, the interest in democratic approaches to assessment has been the common thread among many studies over the last two decades. Brown (1993) shows how test-takers' views can be used in the development of a test for specific language use. Wall et al. (1994) also reflect on students' views in the validation of a placement test. Norton and Starfield (1997) emphasize the accountability in language testing through making the assessment criteria explicit to the test-takers. In the same vein, Yu (2007) attempts to involve the test-takers in the development of a scoring template for the evaluation of their written summaries and by comparing that with those of expert native speakers aims to ameliorate the unfair testing hierarchy between the testers and the test-takers.

When it comes to the rubrics in particular, there have been many studies over the past decade that have emphasized the formative potential of the rubrics by clarifying teacher's expectations, diagnosing the learners' strengths and weaknesses and helping them carry out self-evaluation (Becker, 2016; Ferris & Hedgcock, 2013; Panadero & Johnson, 2013). In the same line, some studies have shown that learners' viewing of a scoring rubric prior to completing a writing task can help them to write better (Howell, 2011; Sundeen, 2014). Similarly, some studies have emphasized that including the students in the co-creating of a rubric helps them to become actively engaged in developing the criteria for their own assessment (Panadero & Johnson, 2013; Skillings & Ferrell, 2000; Sundeen, 2014). What all these studies

share is their emphasis on the active involvement of the students in their assessment in order to make the assessment more ethical, democratic and fair (Lynch, 2001; McNamara, 2001; Shohamy, 2001).

In line with the above, the present study used a pre-post-test research design to explore this issue in a particular Iranian EFL writing assessment context. The aim here was to determine if sharing the rating criteria included in a particular rating scale with the student writers would make any difference to the quality of their writing.

The Iranian EFL writing assessment context has particular features. Results of Ghanbari, Barati and Moinzadeh's (2012) countrywide survey showed that there was a vague situation with regard to writing assessment in the Iranian context. In other words, in the absence of the writers' awareness of the rating criteria, incongruence between the teaching and the assessment of writing occurred. In fact, teacher raters assessed aspects of writing which had not received equal importance in their teaching. Moreover, the existing gap between the teaching and assessment of writing had created an ambiguous construct of writing for them. In other words, they did not know the criteria that their compositions would be rated upon. In addition, teacher-raters usually did not provide the learners with the required explanation on the outcome of their rating. Apparently, the passive role of the test-takers in their own assessment had created a wide gap between the teaching of writing and its assessment in the context. Based on this, the present study attempted to investigate whether involving the learners in their own assessment by asking the learners to view the rubric ahead of their writing task would improve their written texts.

This study

It is clear that the test-takers are greatly disadvantaged in this particular assessment context. Therefore, it seems that the introduction of a rating scale and sharing the rating criteria with the test-takers would enhance their involvement in the rating task and consequently it would affect their writing performance. In order to achieve this goal, the present study addressed the following major research question:

> Does sharing the rating criteria with the student test-takers make any difference to their writing performance?

Method

Participants

94 randomly selected learners (both male and female) participated in this study. The learners' age ranged from 20 to 25. The majority of the learners in this study were undergraduate students who were not satisfied with university English language learning courses and wished to develop a functional competence in all 4 skills including writing. The learners were considered as intermediate learners following the Institute's placement procedure. However, to obtain a homogeneous sample for the present study, an Oxford Placement Test (OPT) was run prior to the study. Finally, 64 learners who were considered as intermediate based on OPT guidelines were assigned into one of two groups: experimental or control.

Context of the Study

The study was conducted at the Zabansara language Institute in Bushehr, a city in southwest Iran. The Institute is well-known in the context for its emphasis on developing the four language skills following the communicative language teaching approach. In addition to general English courses, the Institute

offers separate courses for each of the four language skills. Since productive skills have long been under-estimated in the Iranian EFL context, the Institute aims to encourage the development of all skills including writing. Using up-to-date technological devices and text-books, the Institute also attempts to enrich its language learning courses. All these have turned the Institute into a dynamic language learning center and consequently it has attracted many students who are interested in language learning.

Instruments

Three main instruments were used in this study: the Oxford Placement Test (OPT), a free writing task and the ESL Composition Profile (Jacobs, et al., 1981) as an analytic rating scale. The OPT was used to determine the proficiency level of the learners at the start of the study. The test developed by Allen (2004) is a valid instrument to measure the general language proficiency of the learners. Further, in order to measure the writing ability of the learners, they were asked to write early in the study. In addition, the learners in the experimental group were given the rating scale developed by Jacobs et al. (1981). The scale is an analytic one which measures writing on five distinct components: content, organization, vocabulary, language use and mechanics. In addition to a final aggregate score (out of 100), students receive individual scores for each of the components in the scale.

Data Collection Procedures

Four of the learners (1 in the experimental and 3 in the control group) did not attend regularly so they were excluded in the later analyses. The free writing task acted as a pre-test of students' writing ability. The researcher who was the teacher herself taught writing in both groups. However, the way the learners' essays were treated differed in the two groups. The procedure followed in the experimental group was as follows: when learners were asked to write an essay in the classroom, the rating scale (i.e., ESL Composition Profile by Jacobs, et al. (1981)), upon which their texts would be scored, was explained. The teacher explained different parts of the rating scale to the students. Since the scale was an analytic one, it was taken as a suitable diagnostic tool to be used in an instructional context (Knoch, 2007). Following the teacher's explanations on the scale, a copy of the scale was distributed among the learners prior to their writing. In other words, the student writers were aware of the criteria that their writing would be assessed upon. In addition, the teacher's explanations made different components of the scale clear to the students.

As mentioned, the teacher rated the compositions based on the scale. When the learners' compositions were returned to them, they studied the categories and comments which they had been familiarized with through knowing the scale prior to writing the composition. After returning the compositions to the students, the teacher explained the scale in relation to the way the learners' compositions had been rated. Further, the learners could consult the scale provided to them to better understand the teacher's rating procedure and the scores assigned to them. Some of the learners could not understand some of the teacher's comments; however, with the aid of the scale and the teacher, the ambiguities were clarified. In fact, by using the rating scale the learners could better understand the ratings from the teacher. The same procedure continued for eight weeks with the experimental group.

On the other hand, the learners in the control group experienced the conventional method of writing courses in the Institute. Following writing teaching, the learners were assigned a topic to write about. The teacher rated the texts and assigned scores to the compositions. In fact, the learners were not aware of the criteria that their texts were rated upon. They only received the scores assigned to their texts. Compared with the experimental group, the learners in the control group were left to infer the rating criteria on their own. The teacher used the ESL Composition Profile to score the learners' compositions, however, learners in the control group were not informed about the particular scale used or the rating procedure adopted by the teacher.

Following eight weeks of instruction, learners in both groups were asked to write an argumentative essay similar to the one administered prior to the study. The compositions collected in this way were rated

by the teacher-researcher. Further, in order to establish the reliability of the ratings conducted by the researcher, a researcher's colleague who was an experienced rater was asked to rate a portion of the essays. The analysis of the two raters' ratings showed a considerably high level of reliability between the two sets of ratings ($r = 0.95$, $p < .05$).

Data Analysis

In addition, an independent-samples T-test was run to compare the performance of the experimental and control groups at the end of the study.

Results

In this section, the results of the pre-post-test are explained. Table 1 below shows the descriptive statistics of the control and experimental group prior to the study.

TABLE 1
Descriptive Statistics of Control and Experimental Groups' Writing Score Prior to the Study

	N	Minimum	Maximum	Mean	SD
Experimental	30	30	90	65.53	18.48
Control	30	31	90	64.53	16.48

As Table 1 shows, the performance of the two groups on the writing test was close to each other at the beginning of the study. Table 2 significantly confirms this finding by showing that there was no significant difference between the two groups early in the study.

TABLE 2
Independent Samples T-Test between Control and Experimental Groups' Writing Score Prior to the Study

	Levene's Test for Equality of Variances		t-test for Equality of Means						
								95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Equal variances assumed	.648	.424	.221	58	.826	1.000	4.523	-8.053	10.053
Equal variances not assumed			.221	57.253	.826	1.000	4.523	-8.055	10.055

As Table 2 shows, there were no significant differences between the two groups at the beginning of the study ($t = .221$, $p > .05$). Descriptive statistics related to the performance of the two groups after writing the post-test are shown in Table 3.

TABLE 3
Descriptive Statistics of Control and Experimental Groups on the Writing Post-test

	N	Minimum	Maximum	Mean	SD
Experimental	30	40	94	71.63	16.447
Control	30	33	81	63.37	14.954

Table 3 shows the two groups considerably differed in their mean performance on the post-test. Table 4 shows the results of running an Independent-samples t-test. Table 4 indicates that there was a significant difference between the performances of the two groups on the writing post-test ($t = 2.03$, $p < .05$).

TABLE 4
Independent Samples T-Test between Control and Experimental Groups on the Writing Post-test

	Levene's Test for Equality of Variances		t-test for Equality of Means						
								95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Equal variances assumed	.509	.478	2.037	58	.046	8.267	4.058	.143	16.391
Equal variances not assumed			2.037	57.482	.046	8.267	4.058	.141	16.392

Since the ESL Composition Profile provides separate scores for each of the five components in addition to the one total score, performance of the learners in both groups was also analyzed on each of the components. As an example, Table 5 shows how those learners who had been provided with the scale were considerably more successful compared with the learners who had received the traditional teaching with regard to the content and organization of their writing.

TABLE 5
Performance of Control and Experimental Groups on Content & Organization

	Experimental			Control		
	N	Mean	SD	N	Mean	SD
Content	30	22.47	2.51	30	15.97	2.14
Organization	30	16.70	1.41	30	14.10	1.58

Note. Content and organization had a total score of 30 and 20 in the rating scale, respectively

Figure 1 also indicates that the learners in the experimental group outperformed those in the control group with regard to other components of the scale as well. In fact, the apparent priority of the learners who were involved in the rating of their composition showed that involvement in the rating task and awareness of the rating criteria had noticeable effects on the improvement of the learners' writing in the experimental group.

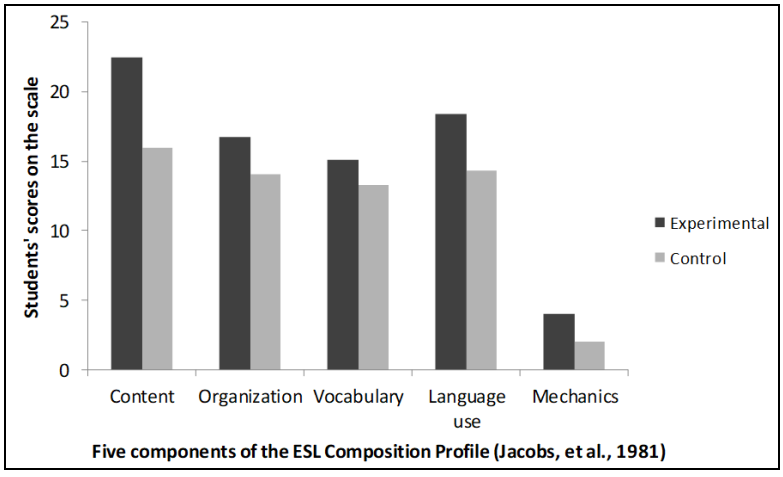

Figure 1. Performance of experimental & control groups on five components of the scale after the study.

Discussion and Conclusion

Findings of the present study provide support for a humanistic approach to assessment. In fact, within the last decades many scholars have emphasized a collaborative approach to assessment in which test-takers' voices as the most important stakeholders in the assessment process should be taken into account. On the other hand, working through a culture of accountability, assessment officials are also in charge of explicitness in the assessment practices. The assumption here is that test-takers should be informed of the assessment criteria and the procedures and processes of assessment should be made explicit to them as well.

In traditional language testing, an instrumental view to the test-takers prevails. This view considers test-takers as passive and neutral instruments who should know specific kinds of knowledge that are defined and shaped by the testing officials. Compliance with the decisions made by testers reinforces the existing unbalanced power relation between the two parties involved in the assessment. It is evident that in this unequal power relationship, it is the tester who leads the whole assessment enterprise including what to test, how to administer the test, how to rate the test, and how to deliver the results. Needless to say, such a mechanical view of test-takers leaves no room for their voice to be heard in the assessment procedure.

Along with the critical movement in language testing, an attempt has been made for more democratic language assessment practices. As an important step, involving the test-takers in the assessment process can redress the unequal power relationship between the test-takers and the testers. Since the pursued target is enhanced learning, sharing the power with the test-takers makes the testing a meaningful activity which in turn promotes the test validity and increases the positive washback on the learning (Bachman & Palmer, 1996; Yu, 2007).

Findings of the present study should be interpreted in light of the above. By involving test-takers in the rating process, they can know the rating criteria and negotiate their perceptions with the teacher-rater. Equally, through establishing a collaborative and participatory atmosphere, test-takers can take responsibility for their learning. The superior writing performance of the learners in the experimental group is evidence in this regard.

A considerable number of studies conducted by many scholars in different ESL/EFL contexts also advocate the outcomes of the present study. The common thread among all these studies is moving towards a more ethical and fair testing practice through active involvement of the test-takers.

One further conclusion that can be made in this study regarding the impact of testing on teaching and learning. Introducing this change in the Iranian EFL writing assessment context can have follow-up effects on the way writing is taught in this context. In other words, when test-takers become involved in their assessment, it paves the way for the teaching of the skill to move towards process-oriented and socio-cultural views (Shrestha & Coffin, 2012; Xi, 2010). The latter considers writing as a social activity, which is constructed through active participation of multiple parties.

Findings of the present study can also have several implications for the teaching and assessment of writing. Some studies conducted on EFL writing assessment have shown that there is an impressionistic approach to rating among the raters (Barkaoui, 2007; Ghanbari et al., 2012). When assessing writing, the majority of the raters draw on their impressions to rate the texts. This practice has unfortunate consequences for the validity of their ratings. In addition, it poses a negative washback effect on the teaching of writing as there is little consistency between the teaching of writing and the subsequent rating procedure. Moreover, student writers do not receive any feedback on the rating criteria. Therefore, reliance on an explicit rating scale would reinforce different components of writing assessment and their expected functions.

Despite the promising findings of the present study, the researcher observed some inconsistencies among the test-takers. Such inconsistent practice was mostly observed at the beginning of the study. The reason can be explained by the unfamiliarity of the Iranian test-takers with this new mode of assessment. In a context where educational measurement still follows the dominant traditional psychometric approach,

involving the learners as active parties in their own assessment would present them with a new mode of practice that they are not accustomed to. In order to provide an insider perspective on the test-takers' attitude and practice in language, further work should investigate this issue qualitatively. The findings of such a study would show to what extent EFL contexts are receptive to such democratic practices. Further, as another line of research, a study can be deemed to investigate whether the learners' involvement in their own assessment can affect their motivation. In fact, awareness of the assessment criteria can provide an explicit route for the learners to further improve their writing. Therefore, in the absence of the strict power hierarchies between the tester and the test-takers which they commonly face within the traditional practices, the test-takers can improve their writing in an informed and confident manner. Hence, this new feeling of achievement can influence their motivation. As another line of research, future studies can investigate whether sharing the assessment criteria with the learners provides any evidence to support the claimed efficiency of explicit corrective feedback through metalinguistic explanations long debated by the scholars (Bitchener & Ferris, 2012).

The Author

Nasim Ghanbari is an assistant professor of applied linguistics at English Language and Literature Department of Persian Gulf University in Bushehr, Iran. Upon joining the Department, she has been teaching writing, testing, and applied linguistics courses. Her area of interest is mainly focused on writing assessment and she seeks interdisciplinary opportunities in this regard.

English Language and Literature Department
Faculty of Literature and Humanities
Persian Gulf University
Bushehr, Iran
Tel: +98 773 122 2321
Mobile: + 98 917 775 3178
Email: btghanbari@gmail.com; btghanbari@pgu.ac.ir

References

Allen, D. *(*2004). The *Oxford Placement Test.* Oxford: Oxford University Press.

Bachman, L.F., & Palmer, A.S. (1996*). Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*(2), 86-107.

Becker, A. (2016). Student-generated scoring rubrics: Examining their formative value for improving ESL students' writing performance. *Assessing Writing*, *29*, 15-24.

Bitchener, J., & Ferris, D. R. (2012). *Written corrective feedback in second language acquisition and writing*. New York: Routledge Taylor and Francis Group.

Brown, A. (1993). The role of test-taker feedback in the test development process: Test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing*, *10*, 277-304.

Ferris, D. R., & Hedgecock, J. S. (2013). *Teaching L2 composition: Purpose, process and practice*. New York, NY: Routledge.

Ghanbari, B., Barati, H., & Moinzadeh, A. (2012). *Development and validation of a local rating scale for Iranian EFL academic writing assessment* (Unpublished doctoral dissertation). University of Isfahan, Iran. Available at https://ganj-old.irandoc.ac.ir

Howell, R. J. (2011). Exploring the impact of grading rubric on academic performance: findings from a quasi-experimental pre-post evaluation. *Journal of Excellence in College Teaching*, *22*, 31-49.

Jacobs, H. L., Zinkgraf, S. A., Wormouth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.

Knoch, U. (2007). *Diagnostic writing assessment: The development and validation of a rating scale* (Doctoral dissertation). University of Auckland, New Zealand. Available at http://researchspace. auckland.ac.nz

Lynch, B. K. (1997). In search of the ethical test. *Language Testing, 14*(3), 315-327.

Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language Testing, 18*(4), 351-372.

McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing, 18*(4), 333-349.

Norton, B., & Starfield, S. (1997). Covert language assessment in academic writing. *Language Testing*, 14(3), 278-294.

Panadero, E., & Johnson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited. *Educational Research Review, 9*, 129-144.

Rea-Dickins, P. (1997). So, why do we need relationships with stakeholders in language testing? A view from UK. *Language Testing, 14*(3), 304-314.

Reddey, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*, 435-448.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language test.* Harlow, England: Longman.

Shrestha, P., & Coffin, C. (2012). Dynamic assessment, tutor mediation and academic writing development. *Assessing Writing, 17*(1), 55–70.

Skillings, M. J., & Ferrell, R. (2000). Student-generated rubrics: Bringing students into the assessment process. *The Reading Teacher, 53*, 452-455.

Sundeen, T. H. (2014). Instructional rubrics: Effects of presentation options on writing quality. *Assessing Writing, 21*, 74-88.

Wall, D., Clapham, C., & Alderson, J. C. (1994). Evaluating a placement test. *Language Testing, 11*, 321-344.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2) 147–170.

Yu, G. (2007). Students' voices in the evaluation of their own written summaries: Empowerment and democracy for test takers. *Language Testing, 24*(4), 539-572.