REDIE. Revista Electrónica de Investigación Educativa
E-ISSN: 1607-4041
redie@uabc.edu.mx
Universidad Autónoma de Baja California
México

González, Elsa Fernanda; Trejo, Nelly Paulina; Roux, Ruth
Assessing EFL University Students' Writing: A Study of Score Reliability
REDIE. Revista Electrónica de Investigación Educativa, vol. 19, núm. 2, abril-junio, 2017, pp. 91-103
Universidad Autónoma de Baja California
Ensenada, México

redie Revista Electrónica de Investigación Educativa
ISSN: 1607-4041
http://redie.uabc.mx

Vol. 19, Núm. 2, 2017

Assessing EFL University Students' Writing: A Study of Score Reliability[1]

Evaluando la escritura de alumnos universitarios de EFL: estudio sobre la confiabilidad de las calificaciones

Elsa Fernanda González (*) e.fernandagonzalez@gmail.com
Nelly Paulina Trejo (*) ntrejo@uat.edu.mx
Ruth Roux (*) rrouxr@uat.edu.mx

(*) Universidad Autónoma de Tamaulipas
(Received: 17 October 2014; Accepted: 17 September 2015)

How to cite: González, E. F., Trejo, N. P. & Roux, R. (2017). Assessing EFL university students' writing: a study of score reliability. *Revista Electrónica de Investigación Educativa, 19*(2), 91-103. https://doi.org/10.24320/redie.2017.19.2.928

Abstract

The assessment of English as a Foreign Language (EFL) writing is a complex activity that is subject to human judgment, which makes it difficult to achieve a fair, accurate and reliable assessment of student writing (Pearson, 2004: 117; Hamp-Lyons, 2003). This study reports on the variability that exists between the analytical grades that 11 Mexican EFL university teachers awarded to five writing samples. It describes the raters' views on writing assessment and their use of analytical scoring rubrics. Data obtained from the grades awarded to each paper, and a background questionnaire, suggested that great variability was found between grades, and raters differed in their levels of leniency and severity, suggesting that having similar backgrounds and using the same rubric are not enough to ensure rater reliability. Participants' perceptions were found to be similar in terms of the use of analytical rubrics.

Palabras: EFL writing, EFL writing assessment, Rater reliability, Analytical scoring rubric.

Resumen

La evaluación de la escritura en inglés como lengua extranjera (EFL) es un proceso que depende del juicio humano, por lo que es difícil de obtener evaluaciones justas, acertadas y confiables (Pearson, 2004, p. 117; Hamp-Lyons, 2003). Este estudio reporta la variabilidad existente entre las calificaciones analíticas que 11 docentes Mexicanos universitarios de EFL proporcionaron a cinco trabajos escritos. Describe las percepciones de los participantes en torno a la evaluación de la escritura y el uso de las rúbricas de evaluación. Los datos obtenidos de las calificaciones de cada trabajo y de un cuestionario escrito revelaron que existe gran variedad entre las calificaciones proporcionadas y que los evaluadores difieren en sus niveles de exigencia, sugiriendo así que antecedentes homogéneos y el uso de una misma rúbrica

[1] The authors presented preliminary results of this study at the 2014 TESL Canada Conference at the University of Regina, Saskatchewan, Canada, May 8-10, 2014

no son suficientes para obtener confiabilidad en las evaluaciones. Las percepciones de los participantes fueron similares en relación al uso de las rúbricas.

Palabras clave: Inglés como lengua extranjera, Evaluación de la escritura, Confiabilidad, Rúbrica de evaluación.

I. Introduction

Edward White (1990) considers that writing assessment largely depends on the "discourse community" in which it takes place, thereby suggesting that writing assessment depends on the people involved and the assessment context. Research has found that many contextual factors besides student performance contribute to grade reliability; in other words, the ability of a test score to be replicable from one test occasion to another (Hamps-Lyon, 2003; Kroll, 1998). For instance, raters differ when they judge students' writing ability depending on their linguistic (Shi, 2001) and educational background (Mendelsohn & Cumming, 1987). Their judgment also depends on the type of scoring scale used and the interpretation that raters give to the scale (Bacha, 2001; Barkaoui, 2007; Knoch, 2009; Saxton, Belanger, & Becker, 2012; Attali, Lewis & Steier, 2012). This study describes the analytical grade variability found between the grades that 11 Mexican university EFL teachers awarded to five sample papers written by university students. It also intends to provide participants' perceptions towards the use and the role of scoring rubrics in the assessment of EFL writing. The following section discusses some of the literature that supports this study.

The assessment of English as a Foreign Language (EFL) or English as a Second Language (ESL) writing is a difficult task that tends to be time consuming, and which teachers may attempt to avoid in their daily practice. Hamp-Lyons (2002) explains that assessment is not value-free and it cannot be separated from the writer's identity and the undeniable effects of washback (p. 182). Researchers have explored these effects for some time, with rater background being one of the most researched factors. For instance, experts have found that raters respond to different aspects of writing, and they do so with some internal inconsistency depending on, for example, their experiential background and their views on the students' linguistic and rhetorical backgrounds (Hamp-Lyons, 1989). Raters also judge students' writing ability differently depending on their academic background and sex (Vann, Lorenz & Meyer, 1991), and the training received (Weigle, 1994). Studies such as Cumming (1990), Eckes (2008), Esfandiari & Myford (2013), González & Roux (2013), Lim (2011), Shi (2001), Shi, Wan, & Wen (2003) and Wiseman (2012), describe how distinct rater backgrounds influence (or do not influence) their rating behavior, actual scores and scoring procedures. Lim (2011), for instance, focused on experienced and inexperienced raters. This longitudinal study considered how participants scored the writing section of a proficiency test over 3 periods of 12-21 months. Data was analyzed using a Rasch model considering rater severity and consistency. The researcher concluded that novice raters, in their particular context, are able to improve as well as maintain grading quality. Additionally, the author reported that inexperienced raters learned how to grade quite quickly and that grading volume and grading quality may somehow be related. A case study by González & Roux (2013) describes the variability in the analytical scores that two Mexican high school EFL teachers awarded to 18 expository essays by 12 high school EFL students. The evidence revealed significant differences between scores despite both raters having similar teaching experience and academic backgrounds. It was found that one rater was more severe and less consistent in her grades than the other, suggesting that rubrics are insufficient to produce reliable assessments. The researchers concluded that rater background is not the only factor that may influence grade variability and considered that raters' rationale of writing and their expectations of student writing are also part of the grading process and the ultimate score given to the paper.

Scoring rubrics are another factor that researchers have found to be influential in writing assessment. Bacha (2001), Barkaoui (2010) and Knoch (2009) have studied how distinct rubrics, mainly analytic vs. holistic rubrics, make a difference in raters' scoring. Khaled Barkaoui (2010) described how rater background and teaching experience, in comparison to the type of grade scale, influence grade variability. Data was obtained from rater think-aloud protocols and the grades awarded to 180 papers written by ESL learners under real test conditions. Participants in the study included 11 novice and 14 experienced raters who scored 24 ESL papers, 12 silently and 12 with the think-aloud technique, using a holistic and an analytical rubric. Results suggested that the type of rubric had more impact on grades

than rater experience. When grading holistically, rater attention focused on the written piece while analytic grading focused on grading scales and criteria. Wiseman (2012) examined the analytical and holistic grades and decision-making behaviors of eight raters of 78 academic papers by ESL students. Data revealed that lenient-to-moderate raters engaged with the text being analyzed as well as the writer of the text, while the severe raters made few comments on the writers' performance, the text, or their own performance as assessors. It was also found that rater background had an impact on raters' use of scoring criteria. Researchers concluded that feedback on grading performance could impact positively on raters' general performance. Most studies on writing assessment focus on ESL and large-scale assessment contexts in English-speaking countries such as the USA, Canada or the United Kingdom. Information on how non-native speaker (NNS) teachers grade EFL writers in non-English-speaking countries is scarce. There is a considerable lack of research from a Latin American perspective, especially in Mexico, where English is an important part of a student's academic life. Additionally, it is our belief that close attention should be paid not only to raters' perceptions of assessment tools such as scoring rubrics but also to EFL teacher classroom assessment practices, because it is their perceptions that may influence their actual assessment practices. This study provides data that could provide insight that would assist in understanding the assessment process in Mexican EFL contexts by answering the following research questions:

1. Are there significant differences between scores given to the same papers by different raters?

2. What are participants' perceptions towards the use and role of assessment tools, such as analytical scoring rubrics, in their assessment of EFL writing?

This study considered 11 Mexican university EFL teachers as the grading participants. Therefore, the results and information described in this study can only be interpreted within the specific context in which they were gathered. According to Hamp-Lyons (1989), research on writing assessment must take a context-embedded approach. Therefore, no generalization of results is intended; rather, this study seeks to provide an understanding of the individuals and the numerous factors that influence the phenomenon under analysis. The following section describes the methods used to collect and analyze data.

II. Methodology

This study adopted a mixed-methods approach (Cresswell, 2013). In other words, quantitative and qualitative principles were used, considering that combining both types of methods enables a better understanding of the problems and situations under analysis (Cresswell & Plano Clark, 2011). A mixed-methods methodology allows for the combination of different types of information in a single study and may provide solutions to issues that could arise when a single method is used (Cresswell et al., 2003, cited in Glowka, 2011; Creswell, 2013). Data for this study was obtained from the five written samples scored by the 11 EFL teachers and an open-closed questionnaire delivered in participants' L1, Spanish. The scores given to papers were analyzed using descriptive statistics (a quantitative method), while the answers given in the questionnaire were analyzed using a qualitative approach.

The raters. Participants in this study answered a background questionnaire that elicited their general background, teaching experience, and perceptions towards writing assessment and analytical scoring rubrics. The questionnaire included eight closed-ended multiple-choice questions and three open-ended questions. By using an instrument that included both types of questions, the study gave participants the opportunity to express their ideas freely, and researchers' data collection and analysis processes were facilitated (Nunan, 1992). Prior to its use with participants, the questionnaire was piloted (Dörnyei, 2003) with a group of English language teachers who were not part of this study, with the aim of obtaining feedback and determining whether its purpose was being fulfilled. Data obtained from the questionnaire revealed that six raters were male, while five were female; the oldest was a 48-year-old male teacher and the youngest a 26-year-old female participant. While some of the participants had ample experience teaching EFL, others were just beginning their teaching career, with between 1 and 17 years of EFL teaching experience. In terms of their professional background, ten instructors had a bachelor's degree and one had a master's degree. In addition to their majors, seven of the participants had an English

teaching certification provided by Cambridge ESOL, such as the Teaching Knowledge Test (TKT) or the In-Service Certificate in English Language Teaching (ICELT). All of the instructors were working in a language center, as part of a public university, in northeastern Mexico, in the state of Tamaulipas. None of the participants had contact with the writers of the samples used for this study, nor were they familiar with the scoring rubric used to grade the sample papers. With the purpose of keeping raters' personal background as homogeneous as possible and in order to eliminate to a greater extent the impact of background on raters' scores, participants were chosen on the following basis: a) they all had the same L1, b) they were all members of teaching staff at the same language center, and c) they all had the same cultural background. Six of the participants reported having received prior assessment training while the rest had not. Table I shows grading participants' backgrounds.

Table I. Raters' Backgrounds

Participant	Age/Gender	Academic background	Years of experience
A12	35/M	BA Teaching Certification	5
B13	26/F	BA	6
C14	36/F	BA /Teaching Certification	17
D22	28/M	BA /Teaching Certification	7
E20	24/F	BA	1
F5	48/M	BA /Teaching Certification	8
G73	26/F	BA /Teaching Certification	8
H16	41/M	BA /Teaching Certification	7
I9	28/M	BA	8
J4	29/M	MA /Teaching Certification	1
K8	25/F	BA	6

The written samples. Instructor participants were provided with five writing samples, each written by EFL lower-intermediate university students enrolled in a Mexican public university. The task required students to write their opinion on a specific statement in 120 to 180 words. This task was part of the activities included in the students' textbook, which they used in their daily English lessons. None of the teacher participants were familiar with the writers; this eliminated any influence that may have had on their scoring performance. Every rater was provided with a copy of each sample and the analytical rubric they were instructed to use, so they could score each sample on the copy given. The following section describes the scoring rubric.

The analytical rubric. An analytical rubric (Appendix A) was provided to raters for their assessment of written samples. This rubric is based on the principle that writing is composed of several different parts (Weigle, 2002), thus allowing teachers to obtain more information from students' writing performance. Additionally, it is considered more useful for EFL/ESL contexts because it enables assessment of more details (Weigle, 2002; Hamp-Lyons, 1991). Researchers adapted the rubric from several sources such as Jacobs, Zinkgraf, Wormuth, Hartfiel & Hughey (1981) and Weir (1990), following the principles outlined in the Common European Framework of Reference for Languages (2002) and the Council of Europe (2009a, 2009b). This rubric assessed five aspects of writing: content, organization, language use, vocabulary use and mechanics/spelling. Each aspect may receive a score of zero (the lowest score) to five (the highest), so 25.00 was the highest score a paper could obtain. Before being used in this study, the rubric was piloted and revised by three experienced EFL professors who were external to the study and gave feedback for its improvement. The purpose of this piloting stage was to ensure the validity and reliability of the study. Grading participants were not familiar with the rubric, nor were they part of the piloting process.

Data collection. Participants were informed of the nature of their participation before data collection took place and signed an informed consent form in which they stated in print their desire to take part in the study. Then each participant was given a folder, which included the background questionnaire, the five writing samples, a form in which they were to record their scores, and the analytical rubric. Participants were given from three to four weeks to complete the scoring process and agreed with the leading researcher on a specific date for papers to be collected, along with the answers to the

background questionnaire.

Data analysis. The analysis of data was carried out in two main phases. In phase one, the answers to the background questionnaire were analyzed in an attempt to understand the grading participants' perceptions towards writing assessment and assessment tools. Answers were grouped into contrasting categories in order to obtain a general perspective on the information. In the second phase, data obtained from the scores for each sample were entered into a commercial statistics software program with the purpose of obtaining descriptive statistics such as the mean (M) and standard deviation (SD), and a *t-test* was performed to compare the M and SD obtained and identify significant differences. The sum of the five analytical scores given to each paper was considered for this statistical analysis. The calculations obtained were then compared between one paper and another and between one teacher and the other to identify important differences and similarities in the data. With the purpose of ensuring the validity of the data obtained, information was discussed among the authors of this study, and then independently with an external expert researcher.

III. Results

Table II lists the scores awarded by each participant to each aspect of the sample papers. To answer research question 1 (RQ1) – "Are there significant differences between scores given to the same papers by different raters?" – eleven raters scored five papers using an analytical scoring rubric adapted by the researchers. Papers were scored independently and returned to the leading author three to four weeks later. A range of different scores was found in the assessment of sample papers. The lowest scores were those awarded by Participants C and G. The lowest score given by Participant C was 6.00 while the highest was 16.00; Participant G gave 5.00 as the lowest score and 16.00 as the highest. In contrast to these raters, Participant I provided a range of higher scores with 17.00 being the lowest and 23.00 the highest. Additionally, the consistency between scores varied greatly. The SD ranged from 2.28 to 7.17.

Table II. Scores awarded by the raters on each task

	Category	Rater A	Rater B	Rater C	Rater D	Rater E	Rater F	Rater G	Rater H	Rater I	Rater J	Rater K
Sample 1	Content	2	2	1	4	2	3	1	3	4	3	3
	Organization	3	2	1	3	1	2	1	2	4	3	1
	Vocabulary	2	1	2	3	1	2	1	2	3	2	2
	Language	2	2	1	3	1	1	1	3	4	2	1
	Mechanics	2	2	1	3	1	2	1	3	2	2	2
	TOTAL	11	9	6	16	6	10	5	13	17	12	9
Sample 2	Content	3	3	4	5	5	4	2	5	5	4	4
	Organization	2	4	2	4	5	4	2	3	4	5	4
	Vocabulary	3	2	2	4	5	5	3	3	5	4	3
	Language	2	2	3	3	5	4	3	4	4	4	3
	Mechanics	2	2	2	4	5	4	2	4	5	3	3
	TOTAL	12	13	13	20	25	21	12	19	23	20	17
Sample 3	Content	5	5	5	5	5	5	4	5	5	4	4
	Organization	5	5	2	5	4	5	3	5	4	5	4
	Vocabulary	4	5	3	4	4	5	4	5	4	4	4
	Language	4	5	3	3	4	4	3	5	5	4	4
	Mechanics	4	3	3	4	4	4	2	5	5	4	4
	TOTAL	22	23	16	21	21	23	16	25	23	21	20
Sample 4	Content	4	5	3	5	5	3	3	3	5	4	4
	Organization	3	5	3	5	5	4	3	4	5	4	4
	Vocabulary	2	3	2	3	3	4	3	4	1	3	3
	Language	2	3	2	4	3	4	3	4	5	3	2
	Mechanics	2	5	2	3	2	3	3	3	4	4	2
	TOTAL	13	21	12	20	18	18	15	18	20	18	15
Sample 5	Content	4	5	4	5	5	3	2	3	5	4	5
	Organization	3	3	2	5	5	2	1	2	5	4	5
	Vocabulary	3	3	1	4	4	4	2	4	2	3	4
	Language	3	4	2	3	3	3	2	3	4	3	4
	Mechanics	4	5	2	5	3	2	3	2	2	3	4
	TOTAL	17	20	11	22	20	14	10	14	18	17	22

This suggests that Participant D was the most consistent in his scoring (SD=2.28) and Participant E the least consistent (SD=7.17). This data can be compared in table III.

Table III. Comparison of Means and Standard Deviations of scores awarded by each rater

Rater	Min Score	Max Score	Mean	SD
A	11.00	22.00	15.0000	4.52769
B	9.00	23.00	17.2000	5.93296
C	6.00	16.00	11.6000	3.64692
D	16.00	22.00	19.8000	2.28035
E	6.00	25.00	18.0000	7.17635
F	10.00	23.00	17.2000	5.26308
G	5.00	16.00	11.6000	4.39318
H	13.00	25.00	17.8000	4.76445
I	17.00	23.00	20.2000	2.77489
J	12.00	21.00	17.8000	3.49285
K	9.00	22.00	16.6000	5.02991

A *t-test* was performed in an attempt to compare the M of the scores awarded by the most lenient participant (I) and the harshest (Participant C). The results obtained from this calculation indicated that

there was a significant difference, producing t=10.58 and p=0.0 (p<.05). Researchers are 95% confident that the difference in means lies between 6.34 and 10.85. Five papers were considered for the scoring process. Data obtained from the scores awarded to each paper revealed that Sample 1 was the paper that received the lowest average score (M=10.36) and Sample 3 the highest (M=21.00). It is equally important to point out that the average SD ranged from 2.82 (Sample 3) to 4.62 (Sample 2), thus indicating that Sample 3 received the most consistent and reliable scores, and Sample 2 the least consistent. Table IV depicts the data described here. As with the data in table III, a *t-test* was used with the purpose of cross-referencing the scores provided by each rater (table III) with the scores given to each paper (table IV).

Table IV. Comparison of Means and Standard Deviations of scores awarded to each paper

Written Sample	Min Score	Max Score	Mean	SD
Sample 1	5.00	17.00	10.3636	3.95658
Sample 2	12.00	25.00	17.7273	4.62798
Sample 3	16.00	25.00	21.0000	2.82843
Sample 4	12.00	21.00	17.0909	2.94803
Sample 5	10.00	22.00	16.9091	4.15823

The M of the scores for the highest-scoring paper was compared with the M of the lowest-scoring sample, which yielded t=11.47 and p=0.0 (p<.05), indicating a significant difference between scores. Researchers are 95% certain that the difference in means lies between 8.57 and 12.70.

In conclusion, and to answer RQ1, the difference found between the analytical scores is quite significant. Although participants were part of the same working environment, had the same L1, graded the same written samples, and were provided with the same scoring rubric, the consistency and variability of scores varied greatly. However, differences were found in their academic training and number of years of teaching experience. Additionally, only six of the participants had received writing assessment training prior to this study. It is our belief that these differences in teaching experience and training influenced the significant differences found between scores and rater variability. Another possible variable to consider is the interpretation and importance that teacher participants attach to scoring rubrics. This variable is described in the following section.

To answer Research Question 2 (RQ2) – "What are participants' perceptions towards the use and role of assessment tools, such as analytical scoring rubrics, in their assessment of EFL writing?" – and to report participants' views on the use and roles of Analytical Rubrics (ARs) in their assessment of EFL writing, raters answered a background questionnaire that included eight closed-ended questions and three open-ended questions. Six answer choices were given for the eight closed-ended questions and were provided in Spanish, the raters' L1. In terms of their use of rubrics, five participants stated that they "always" used rubrics in their regular assessment practices while four claimed to use them "often", and two only used rubrics "sometimes". Eight participants "often" use ARs in their regular practice while one "hardly ever" uses them. One reported using holistic rubrics or a combination of both depending on her teaching purposes. Therefore, it can be noted that most participants use rubrics in their assessment practices on a regular basis, with ARs being preferred to holistic ones. However, it may be inferred that those participants that claimed to use rubrics "often" and "sometimes" during their regular assessment practices are occasionally assessing writing based on their own judgment without the use of any type of assessment tool, resulting in unreliable assessment of their students' written work. It is preferable that every assessment and evaluation process that students undergo be as objective and reliable as possible. In terms of the role of ARs in writing assessment, the questionnaire focused participants' attention on three of the main roles of a rubric: a) ease of assessment, b) objectivity of assessment and c) efficiency of assessment. Regarding the ease of the process, in the data obtained, seven raters considered that using a rubric "always" makes their assessment process easier, four "often", and one stated that using them "sometimes" makes the process easier. Regarding the objectivity of their assessment, nine instructors considered that using a rubric "always" makes their assessment more objective while three chose the option "often". Finally, in terms of efficiency, ten stated that using a rubric "always" makes their

assessment of writing more efficient, whereas two grading participants chose "often". This data allows us to report that most participants consider rubrics to positively influence their assessment and consider them a tool that makes their practice easier, more objective and more efficient. However, their knowledge of these advantages is not enough to obtain reliable assessment practices; more remains to be done. This is further discussed in the following section.

IV. Discussion of Results

The purpose of this study was to reveal if there existed significant differences between the analytical scores that university EFL teacher raters awarded to five writing samples. In addition, it describes grading participants' views on scoring rubrics and their role in writing assessment. Significant differences were found between the scores that the 11 participants gave to the writing samples, even though the same AR was used, and despite many similarities in rater backgrounds. Additionally, it was found that raters differed in their levels of severity and leniency (Participants C and G were harshest vs Participant I as the most lenient) and in the consistency of their scores (Participant D as the most consistent vs. Participant E as the least consistent). These results suggest that similar backgrounds and rubric use are not enough to obtain consistent scores. Other factors such as rater perceptions and previous assessment training experience may also influence writing assessment. These results echo those found by González & Roux (2013), who state that homogenous backgrounds and the use of a rubric are not enough to obtain consistent scores. In their study, scores given by two high school EFL teachers differed greatly as a result of differences in their perceptions of writing and expectations of students' performance. In this study, it is our belief that the variation in scores could have been influenced by participants' use of the rubrics in their regular practice and their views towards the role that analytical rubrics have in writing assessment. Only five raters stated that they always used rubrics in their regular assessments, although the majority considered that using rubrics makes their writing assessment easier, more objective and more efficient. Additionally, it is believed that the raters' assessment training experience played a role in assessment outcomes. Most of the participants (nine raters) had previously received assessment training at different points in their professional careers, which provided these raters with specific knowledge about assessment that the rest did not have. However, the question arises as to the content covered in these training sessions that raters attended. Whether assessment practice is part of training or not is a crucial issue in training teachers to assess writing. By providing teachers with the opportunity to practice and use the assessment tool that the school or institution encourages, assessment reliability may be improved (Weigle, 1994). These results are also in line with those found by Wiseman (2012), who concluded that rater background had an impact on raters' use of scoring criteria. In this study, teaching experience was one factor that varied between grading participants. Teachers had between 1 and 17 years of experience and this may also have had an impact on the variability found between scores. Experienced teachers may rely on different teaching and assessment techniques acquired throughout their years of experience that novice teachers may not be able to use. Wiseman (2012) ends her report by stating that grading participants could benefit from receiving feedback on their grading performance to improve their assessment. It can be inferred from this conclusion that Wiseman is in favor of teacher assessment training as a tool to improve assessment reliability.

On the other hand, the results of this study differ from those found by Barkaoui (2010), in which the use of rubrics was compared alongside teacher participants' level of experience, and the impact that these factors had on score variability was analyzed. It was found that the type of rubric used had a greater impact on rater reliability than years of teaching experience. In the present study, researchers considered that teaching experience could have had an influence on raters' performance and therefore score reliability. Finally, it is important to point out that this study was limited to analyzing inter-rater reliability. However, as reported by Saxton et al. (2012), it may be useful to focus on inter- and intra-rater reliability and examine both types of rater behavior as a path to improving assessment processes. Therefore, future research may focus on comparing the differences found between different raters as well as those found between scores awarded by the same rater. Additionally, future research may examine the impact of assessment training on assessment reliability.

V. Conclusions and Teaching Implications

The results of this study allowed the researchers to conclude that the use of assessment tools, such as scoring rubrics, is not enough to improve rater reliability (González & Roux, 2013; Saxton et al. 2012; Weigle, 1994), and therefore these results may have important implications for teaching practice. First, it is important to consider that rubrics are tools that teachers may use to facilitate their assessment of writing. However, their use does not solve the issue of subjectivity. Unfortunately, writing assessment depends on the judgment of raters, their interpretations of students' work and the rubric in use. Secondly, the results reported in this study may emphasize the importance of teacher training and professionalization. By providing opportunities for professional development, and specifically assessment training or assessment literacy, to teachers within the same institution, rater reliability can be improved and thus students' assessment becomes more valid and reliable. Finally, this small-scale project analyzed information obtained from a small group of teachers working at a language center as part of a public university. Therefore, no generalizations are made. Instead, this study seeks to light the way for those teachers in similar situations or contexts and aid them in the decision-making processes that may allow them to improve their writing assessment practice.

References

Attali, Y., Lewis, W. & Steier, M. (2012). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing, 30*(1), 125-141.

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System, 29*(3), 371-383.

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

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly,7*(1), 54-74.

Council of Europe. (2002). *Common European framework of reference for languages: Learning, teaching and assessment*. Strasbourgh, FR: Author.

Council of Europe. (2009a). *The manual for language test development and examination*. Strasbourgh, FR: Author.

Council of Europe. (2009b). *Manual for relating language examinations to the common european framework of reference for languages: Learning, teaching and assessment*. Strasbourgh, FR: Author.

Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: Sage Publications.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*, 31-51.

Eckes, T. (2008). Rater types in writing performance assessments: a classification approach to rater variability. *Language Testing, 25*(2), 155-185.

Esfandiari, R. & Myford, C. (2013). Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays. *Assessing Writing, 18*(2), 111-131.

Gonzalez, E.F. & Roux, R. (2013). Exploring the variability of Mexican EFL teachers' ratings of high school students' writing ability. *Argentinian Journal of Applied Linguistics, 1*(2), 61-78.

Glówka, D. (2011). Mix? Yes, but how? Mixed methods research illustrated. In M. Pawlak (Ed.), *Extending the boundaries of research on second language learning and teaching* (pp. 289-300). Poland: Springer.

Hamp-Lyons, L. (1989). Raters respond to rhetoric in writing. In H. Dechert & C. Raupach (Eds.), *Interlingual processes* (pp. 229-244). Tubingen: Gunter Narr Verlag.

Hamp-Lyons, L. (1991). *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex Publishing Corporation Assessing.

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing, 8*, 5-16.

Hamp-Lyons, L. (2003) Writing teachers as assessors of writing. In Kroll, B. (Ed.) *Exploring the dynamics of second language writing* (pp.162-189). New York: Cambridge University Press.

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

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275-304.

Kroll, B. (1998). Assessing writing abilities. *Annual Review of Applied Linguistics,18*, 219-240.

Lim, G. (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.

Mendelsohn, D. & Cumming, A. (1987). Professors' ratings of language use and rhetorical organizations in ESL compositions. *TESL Canada Journal, 5*(1), 9-26.

Nunan, D. (1992). *Research methods in language learning*. New York: Cambridge University Press.

Pearson, P.C. (2004). Controversies in second language writing: Dilemmas and decisions in research and instruction. The University of Michigan Press.

Saxton, E., Belanger, S. & Becker, W. (2012) The Critical Thinking Analytic Rubric (CTAR): Investigating intra-rater and inter-rater reliability of a scoring mechanism for critical thinking performance assessments. *Assessing Writing, 17*(4), 251-270.

Shi, L. (2001). Native and non-native speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18*(3), 303-325.

Shi, L., Wan, W. & Wen, Q. (2003). Teaching experience and evaluation of second language students' writing. *The Canadian Journal of Applied Linguistics, 6*, 219-236.

Vann, R., Lorenz, F. & Meyer, D. (1991). Error gravity: Faculty response to errors in the written discourse of non-native speakers of English. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 181-195). Norwood, NJ: Ablex.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*, 97-223.

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

Weir, C. J. (1990). *Communicative language testing*. NJ: Prentice Hall Regents.

White, E. M. (1990). Language and reality in writing assessment. *College Composition and Communication, 41*(2), 87-200.

Wiseman, C. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing, 17*, 150-173.

Appendix A: Analytical rubric

Score	Content	Organization	Use of Language	Use of Vocabulary	Mechanics and Spelling.
5	Text shows knowledge of the topic and gives details or examples to support main ideas. Text fully corresponds to task requirements. Communication is effective.	Organizational skills are present in the text making flow and coherence of ideas smooth. Main ideas and structure of text are easily found and logically sequenced.	Text makes use and maintains use of complex language structures effectively. There are no errors of idioms, collocations and grammar in general. Facility in use of language is apparent.	Demonstrates sophisticated and broad use of vocabulary. Effective and appropriate use of idiomatic expressions and colloquialisms; shows awareness of connotations and their meaning.	Writing shows mastery of punctuation and spelling conventions. Capitalization and paragraphing errors and typos are not found.
4	Task is answered in its majority but information may be redundant or unnecessary. Some detail is given. Sufficient development of main ideas. Some gaps may be found among information.	Adequately organized with the use of organizational patterns and connectors but sequencing of information is incomplete. Connection of main ideas may be lost but meaning is still understood.	Grammatical accuracy consistently maintained. Few errors of idioms, collocations and grammar in general. Complex sentences present minor errors.	Demonstrates sophisticated use of vocabulary. Good command of idiomatic expressions and colloquialisms. Minor errors in vocabulary use.	Writing shows occasional errors of punctuation and spelling conventions. Capitalization and paragraphing errors and typos are occasionally found.
3	Task is addressed adequately but information may be missing. Some details are used to support the main idea. Shows some knowledge of the main topic and limited development of main ideas.	Some organizational skills are present. Use of cohesive devices makes text clear and understood. Occasional deficiencies can lead to "jumpiness" among information.	Some grammatical "slips" may be found. Grammatical errors such as verb tense, verb agreement, number, word order, articles, pronouns, and prepositions are found but they do not lead to misunderstanding. Context given in text allows for interpretation of meaning.	Vocabulary accuracy is high though occasional errors may be found. Adequate and appropriate word/idiom choice and use. Some incorrect word choice does occur without impeding communication.	Writing shows few errors of punctuation and spelling conventions. Few capitalization and paragraphing errors and typos are found.
2	Task reveals little relevance to the topic. Major gaps in information are found and insufficient details to support main ideas are given. Inappropriate information. Pointless repetition of information.	Small pieces of text are linked with basic connectors. Unsatisfactory cohesion may cause most but not all of the information to seem sloppy and non-fluent.	Frequent grammatical inaccuracies found. Frequent and basic errors of tense, agreement, number, word order, articles, pronouns, and prepositions are found. Understanding of ideas is seldom confusing.	Sufficient control of elementary vocabulary to express basic ideas. Repetition of vocabulary is frequent. Frequent misuse of word form use, word/idiom choice and use, making communication confusing.	Writing shows frequent errors of punctuation and spelling conventions. Capitalization and paragraphing errors and typos are frequently found. Meaning may be confusing.

1	Task presents limited relevance to main topic. Inadequate development of topic. Details are not given.	Groups of words connected with simple connectors such as "and", "but" or "because". Cohesion is almost absent. Connection among ideas is difficult to find making information confusing or misleading.	Almost all or most of the basic grammatical constructions are inaccurate. Major issues in simple sentences. Errors of negation, agreement, number, word order, articles, pronouns, and prepositions frequently found. Understanding of information difficult.	Text has little knowledge of English vocabulary, idioms and word forms. Language sufficient for coping with simple survival needs. Information is basically translated. Inappropriate choice of word forms.	Almost all spelling is inaccurate and the text shows an ignorance of punctuation conventions . Text is dominated by capitalization and paragraphing errors and typos. Meaning is obscured.
0	Task does not reveal development topic. Totally inadequate answer to task. No details are given. Content insufficient to assess.	Cohesion is totally absent. Writing is fragmented making communication impossible to obtain. Lack of structure in information leads to absence of organization. Content insufficient to assess.	All language use is inaccurate. Meaning obscured. Content insufficient to assess.	No apparent vocabulary use or vocabulary comprehension is present in text. Content insufficient to assess.	All spelling is inaccurate and the text shows an ignorance of punctuation conventions . Text is dominated by capitalization and paragraphing errors and typos. Meaning is obscured. Content insufficient to assess.

Adapted from: Council of Europe, 2009; CEFR, 2002; Jacobs et al., 1981; Weir, 1990.