Argentinian Journal of Applied Linguistics
Vol. 1, No. 2, November 2013, 61-78

FAAPI

Exploring the variability of Mexican EFL teachers' ratings of high school students' writing ability

Elsa Fernanda González Quintero*
Universidad Autónoma de Tamaulipas, Mexico

Ruth Roux Rodríguez
Universidad Autónoma de Tamaulipas, Mexico

(Received 28/04/13; final version received 12/8/13)

Abstract

EFL teachers generally use scoring rubrics to assess students' writing seeking to rate consistently and objectively so that assigned scores reflect students' abilities rather than other unrelated factors. Research has found that raters' judgment varies, depending on their linguistic, educational and professional/personal background. This study investigated the degree of difference among the scores of two Mexican raters of academic papers written in English as a foreign language by six high school students. Data came from two sources: 18 scored papers and raters' comments during a semi structured interview. Results indicated that although the raters' backgrounds were very similar, their judgments differed as a result of their personal perceptions of writing, scoring rubrics and writing assessment.

Keywords: EFL writing; rater variability; rater background; foreign language writing assessment; writing assessment.

Resumen

Los docentes de inglés como lengua extranjera generalmente utilizan rubricas para evaluar la escritura de sus estudiantes de manera objetiva y consistente. Estudios han encontrado que los juicios de los evaluadores varían dependiendo de sus antecedentes lingüísticos, educativos, y profesionales. El presente estudio investigó si existían diferencias en los puntajes de dos evaluadores mexicanos de textos escritos en inglés por seis estudiantes mexicanos de bachillerato. Los datos se obtuvieron de dos fuentes: los 18 ensayos evaluados y una entrevista semi estructurada con los evaluadores. Los resultados indicaron que, aun cuando los antecedentes de los evaluadores eran muy similares, sus juicios acerca de las habilidades de los escritores diferían como resultado de la diferencia de sus percepciones personales sobre la escritura, las rubricas y la evaluación de la escritura.

Palabras clave: Escritura en inglés como lengua extranjera, variabilidad del evaluador antecedentes del evaluador, evaluación de escritura en lengua extranjera, evaluación de escritura.

* Corresponding author, e-mail: e.fernandagonzalez@gmail.com ISSN 2314-3576 (online)

Teachers of English as a foreign language are increasingly using different kinds of rating scales to assess students' writing. Rating scales are thought to facilitate teachers' interpretation of the quality of students' writing performance, especially after extensive training (Weigle, 1994). It is desirable that raters rate consistently and objectively so that their ratings reflect students' ability. Rating, however, is a complex and error prone process. Research has found that raters' judgment of the same texts often varies. A number of studies have investigated this variable behavior of raters through the use of the Rasch statistical model (Engelhard, 1992; Kondo-Brown, 2002; Mayford & Wolfe, 2004). The model, however, is not only difficult to interpret and use by EFL teachers; it is intended for large-scale testing situations.

Other studies have used alternative statistical analysis combined with data from questionnaires and interviews. Hamp-Lyons (1989), for example, has found that writing assessment is affected by distinct factors, such as the nature of the writing task, the scoring procedures, and the characteristics of the raters. Raters also differ when they judge the writing ability of students depending on their linguistic (Shi, 2001) and educational background (Mendelsohn & Cumming, 1987).

This study focuses on the ways in which the raters' background influences their judgment of the writing ability of EFL student writers. Specifically, it examines the ways in which two raters score the same set of papers written by EFL high school students and the rationale of their judgments. The following section presents a review of the literature on the relationship between raters' background and their rating behavior.

Literature review

The assessment of English as a Foreign Language (EFL) writing is a field that seeks objectivity, validity and reliability among its practitioners for the benefit of language students and for language program development. It is an area that makes an impact on student and teachers´ lives because many life-changing decisions may be made depending on assessment results. As Hamp-Lyons (2002) explains assessment is not value-free and it cannot be separated from who the writer is and from the undeniable effects of "washback" (p. 182) on teaching and learning. However, writing assessment faces difficulties in attaining the validity and reliability needed because scoring procedures will always be subject to human judgement therefore making fair and accurate assessment of student writing difficult to actually reach (Pearson, 2004, p. 117).

> …If raters A and B disagree on how to rate an essay, how can the final score (e.g., an average score or a total) be fair or meaningful to the writer? Similarly, if a rater scores one way when fresh and another when fatigued, how can a student whose paper is read when the reader is tired be rated

> fairly? (Pearson, 2004, p.124).

Therefore the variability among raters judgment is a matter of concern in EFL writing instruction. Scoring rubrics have been studied to find out if raters using them differ in the levels of leniency or severity (McNamara, 1996). Raters respond to different aspects of writing and they do so with some internal consistency 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; Cushing, 1994).

In a recent study conducted by Wiseman (2012) 8 raters were examined as well as their decision making behaviors when rating 78 academic papers. Their ratings were analyzed to describe the degree to which raters were consistent in their ratings of English as a Second Language (ESL) writing by using a mixed methods approach: quantitative and qualitative analysis. The author concludes that

> think aloud protocols showed that depending on their background, individual raters engaged with the text or prompt type. Rater background, which might be regarded as another facet of ego, seemed to contribute to raters' expectations of criteria for narrative vs. persuasive essays (Ibid, p. 169).

Rasch Analysis suggested that moderate-lenient raters differed from those stricter in their focus of attention when rating. Severe raters tended to concentrate on the performance descriptors rather than engaging with the text or the writer. It is also concluded that raters may also benefit from performance feedback as a means to analyze rating performance for any necessary adjustments.

Another factor that seems to influence raters' judgment is their linguistic background. Studies conducted in English speaking countries have compared the scoring behaviors of non-native speaker (NNS) and native speaker (NS) raters. Findings of these studies are mixed. In some cases NNS were more severe than NSs (Santos, 1988). In other cases NNS were more lenient than NS in different aspects of writing (Brown, 1995). These results could exemplify how distinct background traits corresponding to each rater can be part of writing assessment variation.

Professional experience is another variable that influences the judgment of raters. This variable includes the educational experience of the raters, their previous teaching experience, and the level of assessment experience, among others. In relation to educational experience, studies have found that the raters' academic discipline impacts

their ratings of EFL students' writing (Mendelsohn & Cumming, 1987; Santos, 1988). Mendelsohn and Cumming (1987), for example, examined the differences between the ratings of engineering professors and ESL professors. Their findings indicated that engineering professors attributed more importance to language use than to rhetorical organization in rating the effectiveness of ESL papers. ESL professors, on the other hand, attributed more importance to rhetorical organization. In another study, Santos (1988) investigated the scorings of 178 non-ESL professors of two ESL students' written compositions. These professors were in the fields of humanities and social sciences (96) and physical sciences (82). Results indicated that the physical science professors were more severe in their scores than the humanities and social science professors.

Raters' perceptions of writers are variables approached by Johnson and Van Brackle (2012) in a study in which raters holistically assessed 358 essays. Raters of these papers were required to take a training assessment course and focus on errors made by African American English (AAE), English as a Second Language (ESL) and standard American English (SAE) writers. Raters were assessors of the Regent Writing Exam at the University of Georgia and therefore part of a large-scale assessment practice. Writers were sophomore students who were required to obtain a passing grade on this writing test to continue their university studies. Each paper received 65-72 ratings. Data found indicated that the 3 different types of writers received distinct ratings and were the AAE who received stricter ratings of errors. Researchers conclude that linguistic discrimination may be present in raters' scores and that raters may find themselves "annoyed" (Ibid, p.46) by the careless errors of AAE writers.

The years of L2 writing instruction experience have also been found to have an influence on raters' scorings. In a study conducted by Shi, Wang and Wen (2003), 46 English teachers from three universities in China were asked to evaluate ten essays written by English majors, and to justify their scores for each essay with qualitative comments. Findings indicated that the most experienced writing teachers gave significantly lower scores than the teachers with less experience in 4 out of 10 essays. Analysis of the qualitative comments on the 4 essays suggested that the experienced teachers made more negative comments on organization, language fluency, ideas and general language. In a different study, Khaled Barkaoui (2010) described how rater background and teaching experience in comparison to the type of rating scale influence rating variability. Participants in the study included 11 novice and 14 experienced raters who scored 12 ESL essays with a holistic and analytic rubric. Results suggested that type of rubric had more impact on rating than rater experience. When rating holistically, rater attention focused on written piece while analytic rating focused on rating scales and criteria.

From a different perspective and with a different research purpose, other studies

have focused on comparing inter- and intra-rater scoring and how their processes vary. Considering that inter-rater variability focuses on how scores vary from one rater to the other and that intra-rater variability describes how a single rater can vary scores on a single paper, Saxton, Belanger and Becker (2012) describe the inter- and intra-rater reliability when using an analytic rubric that focused on writers´ critical thinking skills. Two female raters, who shared equal background information and teaching experience, took part in a rating protocol for this study. Data revealed that both raters were consistent in their intra–rater scoring performance and showed that raters reached acceptable inter-rater reliability when using the rubric. Researchers concluded that training could help raters attain consistency when using a scoring rubric.

Most of the studies on assessment of writing have focused on ESL students in English speaking countries. Information on how NNS rate EFL writers in non-English speaking countries is scarce. Additionally, more information on raters' perceptions and personal use of a rubric is necessary in EFL assessment. The present study investigates the judgments of two Mexican EFL teachers rating the papers of six high school students who were not their pupils. The study aims to explore the raters' rationale for ways in which they judge the students' writing abilities. The research questions to be responded by the study were:

1. Are there significant differences between the scores assigned by raters to the same papers?
2. What are the raters' views on writing ability?
3. What are the raters' views on the use of scoring rubrics to assess students' writing?

Hamp-Lyons (1989) suggested that, as in other human endeavors, research on writing assessment must use a context embedded approach. In this study, therefore, we assumed that raters' views and behaviors could only be interpreted in the context of the specific situation in which they were involved. No attempt was made to measure the effect of a factor to obtain generalizations, but rather, to allow an intensive view of individuals and the many factors that influenced their behaviors.

Methodology

This study used a mixed methods approach with the central premise that the use of quantitative and qualitative approaches, in combination, provides a better understanding of research problems than either approach alone (Cresswell & Plano Clark, 2011). According to Cresswell (2013), this is a relatively new method that allows researchers to combine distinct types of data and fill in the gaps that one type or another type may

have. Experts have described the concept of the mixed methods approach as one that allows for qualitative and quantitative data to be collected in a single study sequentially and concurrently allowing for different scientific inquiry to be approached (Cresswell et al., 2003 cited in Glówka, 2011). This study collected data through scored essays and a semi structured interview to raters. Essays were scored using statistical analysis methods (quantitative methods) while interview transcripts were analyzed with a qualitative approach. This mixture of methods allowed for data in this study to be analyzed from both perspectives and cross-reference information found. The following sections explain the details of the approach used.

The Raters

Raters were two EFL teachers, one male and one female, who were selected from 6 EFL teachers of a small-size, private high school in northeast Mexico. They were selected because of their similarity in cultural background, mother tongue and professional experience. In the literature review these dimensions of raters' background are considered important sources of variability in writing ability judgment. We chose the participants in such a way as to minimize variability. Although they were part of the teaching staff of their school, neither of them were the writers' instructors at the time of the study. The raters' names were replaced with codes to preserve their anonymity.

To obtain background information from the raters, a background questionnaire was answered in Spanish at the beginning of the study. Rater A was a 37-year old female teacher who had 4 years of teaching experience and a BA in Communication Sciences, a Diploma in Marketing and a Diploma in Teaching Competencies. She also held an In-service-Certificate of English Language Teaching (ICELT).

Rater B was a male teacher of 31 years of age and 5 years of teaching experience. He had a BA in Communication and Public Relations, a Diploma in Teaching Competencies, a Band 1 Certificate in the Cambridge Teaching Knowledge Test (TKT) and the Cambridge In-service Certificate for English Language Teaching (ICELT).

Both Rater A and Rater B had experience working with high school students. Also, they both considered writing an important skill to develop in language learners. Rater A considered that writing provides students with the opportunity to use verb tenses, vocabulary and other language forms in a more formal way. She held a form-focused view of writing as an activity that allows students to practice a specific language form previously learned in class. Rater B, on the other hand, stated that writing is important because it is the visual way in which language is represented. He considered that writing should be done in an organized and coherent way. This instructor viewed writing as an opportunity to communicate rather than to practice language forms.

The raters did not receive any kind of preparation or training prior to the study. They

accepted to blindly rate 18 papers written by 6 high school students during an advanced general English course, taught by one of the researchers. They relied on their judgment and experience to assess the papers, using a predetermined analytic scoring rubric.

The Writing Tasks

Student writers participated in three different writing tasks as part of their English language class. For each task, they were required to develop a descriptive essay of 150-180 words (see writing prompts on Appendix A). Students had opportunities to engage in classroom discussions and brainstorming activities, prior to the writing task. Writing activities took place in the school language laboratory with Internet access. Once students had brainstormed their ideas and had prepared an outline, they developed their papers on a previously created personal weblog. If the writing task was not finished during class time, they were required to finish the composition at home and to upload their work.

Scoring Rubric

Raters were provided with an analytic scoring rubric (Cushing, 2002) to assess several aspects of students' writing. The rubric was adapted from Jacob's et al. (1981, cited in Cushing, 2002) and focuses on 5 aspects of writing: content, organization, vocabulary, language use, and mechanics (see rubric on Appendix B). Each aspect was graded using a 1-6 scale in which 1 indicated the lowest and 6 the highest performance. The scoring rubric was familiar to the raters who had previously used it to assess their students' writing.

Data Collection Procedures

Before initiating the data collection activities, raters were given information on the purpose and procedures of the study. Then they were asked to sign an informed consent and to fill-in a background questionnaire, both written in Spanish. The raters were given a separate, copied set of 18 papers to grade, a copy of the writing prompts given to students, and a copy of the 6-point analytic scoring rubric. They were given two weeks to independently score the papers. Scores were directly written on each paper. Neither names nor any other personal background information of the writers was shared with the raters.

To explore the influence of raters' factors on their judgments, raters were interviewed by one of the researchers, once the scoring was finished. Interviews were conducted in Spanish and lasted approximately 15-25 minutes. Interviews followed a semi-structured format. Semi-structured interviews allow for more flexibility in terms of question structure and follow-up responses of the interviewee (McDonough & McDonough,

1997; Nunan, 1992) which might allow for more interview engagement between the interviewer and interviewee.

Data Analysis Procedures

Analysis of data involved four stages. First, descriptive statistical analyses (the mean and standard deviation) as well as paired samples t-tests for the writing scores given by both raters were conducted. These statistical analyses were calculated to compare the score means and standard deviations of the two raters and determine if there were significant differences between the scores given by Rater A in comparison to Rater B for the 18 papers written by the high school students. In a second stage of analysis, we entered the scoring data into a commercial spreadsheet to create a line-chart that allowed visual representation and analysis. The third stage consisted in identifying aspects that could explain the contrasts in the scores given by the raters to each of the writers for the three writing tasks to design a set of interview questions. Finally, we examined the responses given by the raters to the interview questions to identify aspects of their background that influenced their judgment and scoring behavior. To ensure the validity and reliability of the interview, the researchers analyzed transcripts independently to identify themes then results were cross-referenced and agreed upon.

Space intentionally left blank

Results

Table 1 shows the scores assigned by the two raters to each paper written for the three writing tasks. Only the total scores were considered in the statistical analyses.

		Category	Writer A	Writer B	Writer C	Writer D	Writer E	Writer F
Rater A	**Task 1**	Content	6	4	4	6	4	4
		Organization	6	6	4	6	4	5
		Vocabulary	6	5	5	5	5	6
		Language	6	5	5	6	6	6
		Mechanics	6	5	4	6	5	6
		TOTAL	**30**	**25**	**22**	**29**	**24**	**27**
	Task 2	Content	6	6	6	6	6	4
		Organization	6	6	6	6	6	4
		Vocabulary	6	6	6	4	6	4
		Language	5	6	6	4	6	4
		Mechanics	5	6	5	4	6	4
		TOTAL	**28**	**30**	**29**	**24**	**30**	**20**
	Task 3	Content	6	5	6	6	6	1
		Organization	6	6	6	6	6	1
		Vocabulary	6	6	6	6	6	1
		Language	5	6	6	6	6	1
		Mechanics	6	5	6	6	6	1
		TOTAL	**29**	**28**	**30**	**30**	**30**	**5**
Rater B	**Task 1**	Content	4	4	4	4	4	4
		Organization	5	4	4	5	4	3
		Vocabulary	4	4	4	6	4	3
		Language	5	4	3	6	4	3
		Mechanics	4	4	3	6	4	3
		TOTAL	**22**	**20**	**18**	**27**	**20**	**16**
	Task 2	Content	4	4	4	6	5	2
		Organization	5	4	4	6	5	1
		Vocabulary	4	5	4	6	4	3
		Language	4	5	4	6	4	3
		Mechanics	5	4	4	6	4	4
		TOTAL	**22**	**22**	**20**	**30**	**22**	**13**
	Task 3	Content	5	4	4	5	4	4
		Organization	4	4	4	5	4	4
		Vocabulary	4	4	4	4	4	4
		Language	4	4	4	5	4	3
		Mechanics	4	4	1	4	4	4
		TOTAL	**21**	**20**	**17**	**23**	**20**	**19**

Table 1. Scores assigned by the raters on each task

Are there significant differences between the scores assigned by raters to the same papers?

Table two summarises and compares the means and standard deviations of the ratings of

Rater A and Rater B on the three papers for each of the six EFL student writers. Means ranged from 16.00 to 29.00. Rater B provided lower scores (Group M score of 20.70 vs. 26.11) than Rater A. Although both raters judged that the papers written by Fiona (M of 16.00 Rater B vs. 17.33 from Rater A) had more flaws, they differed in their judgment of the papers with fewer weaknesses. While Rater A assigned the highest scores to the papers written by Albert (M of 29.00), Rater B assigned the highest scores to the papers written by Daniel (M of 26.66). Equally important is to point out that Rater B was more consistent in his scoring. While he obtained a Sd. of 3.71, Rater A was less consistent, obtaining 6.10 as a Sd.

Writer	Rater A		Rater B	
	M	Sd	M	Sd
Albert	29.00	1.00	21.66	0.57
Beatriz	28.66	1.15	20.66	1.15
Charlie	27.00	4.35	18.33	1.52
Daniel	27.66	3.21	26.66	3.51
Emma	28.00	3.46	20.66	1.15
Fiona	17.33	11.23	16.00	3.00
Group M	26.11	6.10	20.70	3.71

Table 2. Comparisons of the scores assigned by Rater A and Rater B

Scoring inconsistency of Rater A is visually represented in Figure 1. Inconsistency in the judgment of the writing ability was particularly evident in the scores assigned to the papers written by Fiona, which ranged from 5 to 27 points.

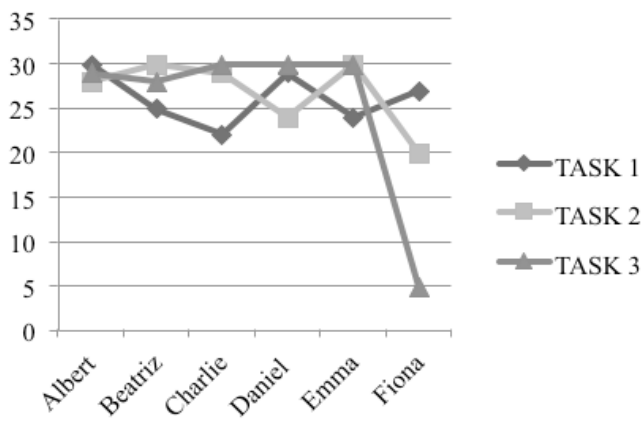

Figure 1. Score results of Rater A

Rater B was more consistent in his judgments of the students' writing abilities as is shown on Figure 2.

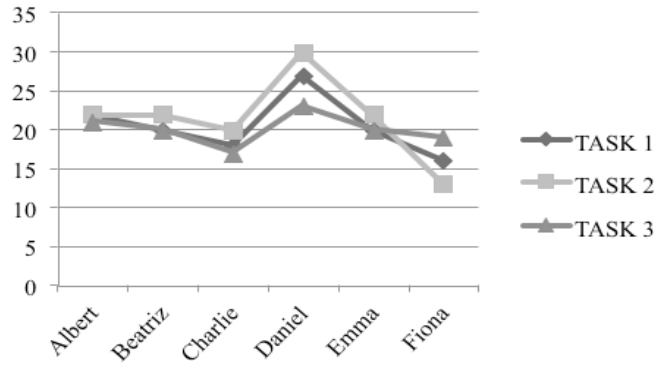

Figure 2. Score results of Rater B

To compare the means of the scores assigned by the two raters, a t-test was used. The t-test assesses whether the means of two groups of scores are statistically different from each other. Results indicated that the scoring of Rater A is significantly different from the scoring of Rater B, $t(34) = 3.2109$, $p < .05$. The researchers are 95% confident that the mean difference lies between 0.8214 and 10.0675. These results suggest that the raters' assessment varied greatly even when given the same written samples and the same scoring rubric.

On Task 1, Rater A gave higher scores on every aspect in comparison to Rater B. Both Raters followed the same descending and ascending patterns of scores on every student, except on Fiona who received 27 points from Rater A and 16 points from Rater B.

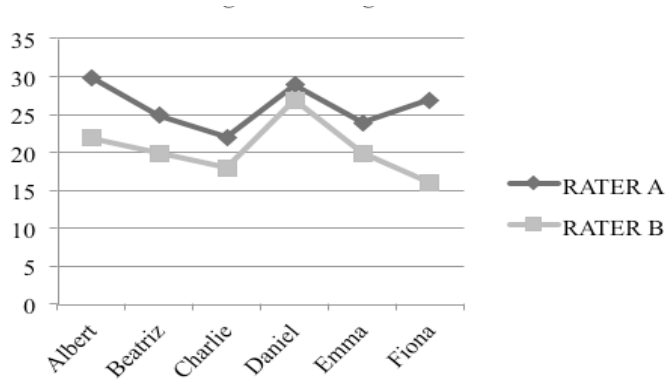

Figure 3. Ratings for Task 1

On Task 2, Rater A (20 points being the lowest) provided higher scores to every student than those given by Rater B (16 points being the lowest score). Figure 4 depicts the differences in scores given to Daniel. Rater B gave a higher score (30 points) to Daniel than Rater A (24 points).

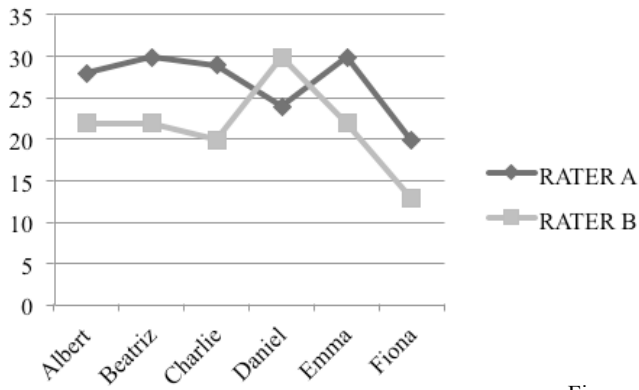

Figure 4. Ratings for Task 2

For Task 3, Raters disagreed again in their scoring. Rater A provided higher scores on all of the papers except the ones written by Fiona. In this case, Rater B assigned 19 points to her paper while Rater A assigned 5 points.

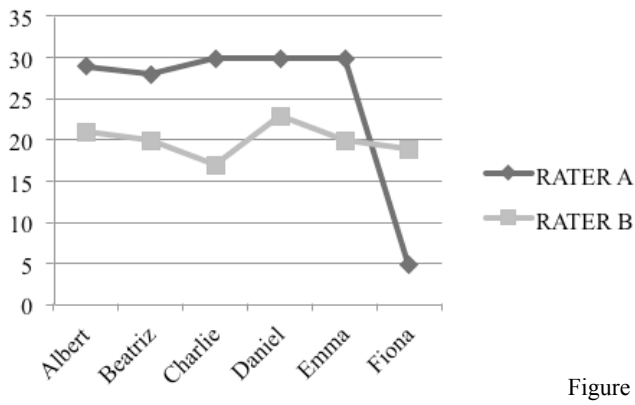

Figure 5. Ratings for Task 3

In sum, the response to the first research question is that there were significant differences between the scores assigned by Rater A and Rater B to the same papers. The raters' background -in terms of age, teaching experience, and educational experience- were similar. They were also part of the same teaching staff in the same private high school. However, differences were found among their views of what writing ability is and on the concept, purpose and use of scoring rubrics to assess writing. It is the researchers´ belief that these differences in writing assessment and rubric use perceptions could have influenced rater variability and rating scores. These perceptions are further discussed in the following section.

How do raters define writing ability?

Rater A was a Mexican female with five years of EFL teaching experience. During the

interview she indicated that she viewed writing ability as an essential part of foreign language competence. She considered that *"language development is incomplete without writing."* Furthermore, Rater A complained about the scarce attention given by other teachers to developing their own and their students' writing ability. The following is an excerpt of her comments.

> Teachers do not teach their students to write and they themselves do not know how to write. We are at a disadvantage because we do not know how to write. It is important for our students to know how to develop a composition and what language structures to use.

Rater A felt great responsibility for teaching students how to write correctly in English and she considered that high school students need to learn how to write distinct types of texts -from letters and emails to essays-, according to their level of English language proficiency. In general terms, Rater A associated writing ability with the correct use of grammar and correct use of each writing genre.

Rater B, on the other hand, was a Mexican male teacher with four years of EFL teaching experience. The interview transcripts reveled that he considered writing as "a more structured way of thinking" and that, to write properly, students need to be aware of the purpose of the text and the target audience. Writing ability, according to his views, takes a long time to develop. He attributed more importance to the ideas that student writers try to express, even when they still have limited grammatical competence. Rater B associated writing ability with the capability of communicating ideas to a particular audience.

In spite of the similarity of their professional backgrounds, native language and years of teaching experience, the raters' views on what writing ability is were very different. While Rater A associated writing ability to the correct use of language structures, Rater B viewed writing ability as the capacity to communicate, in the process of acquiring the foreign language.

When asked what they thought of the writing ability of the specific writers they assessed, Rater A responded that students had good use of language structure and vocabulary, although they did not organize their work properly and did not consider the target audience they were writing for. She added that some did not show they had planned before writing. In spite of the weaknesses identified in the students' papers, Rater A was lenient compared with Rater B. The reason could be that students were demonstrating good use of grammar, the most important trait of writing ability, from her point of view.

Rater B found that the students he assessed for the study did not meet the needs

of their audience because they had not written for an audience in the first place. In his opinion, students had written for themselves rather than for a reader. Rater B gave lower scores than Rater A, probably because he was expecting not only grammatical competence but also sociolinguistic and discourse competence.

What are the raters' views on the use of scoring rubrics to assess the students' writing?

During the interview the raters were also asked about their views on the use of scoring rubrics. Rater A reported that she regularly used rubrics to grade her students' writing. She considered rubrics as easy-to-use tools that make the scoring procedure more objective by allowing the rater to set aside students' variables. The following is a segment of her comments.

> A rubric is a guide that can help you assign a grade because we can tend to be subjective depending on whom we are grading. The rubric allows you to focus on the actual writing, rather on the students.

Rater A indicated that, when scoring the essays for the study, she gave more importance to the correct use of the grammar. This comment seemed contradictory to the use of a rubric in which various components or traits of writing are included.

Rater B commented that he was familiar with the use of rubrics for different purposes and that he used them differently, depending on the class. If in a lesson he emphasized a specific language structure, then he would focus on the use of that structure when scoring. If during lessons organization, content and other text-level aspects were reviewed during lessons, then his scoring would focus on those aspects. Rubrics, in his opinion "…are tools for objective assessment and they facilitate your job when assessing writing because you have specific parameters to follow. It is more objective than just giving a grade based on what I consider good."

Rater B also affirmed that the use of rubrics encourages fair grades for students because it eliminates the influence of the students' personal traits. While grading essays for this study, he found that writers had not written for an audience and he suggested including the aspect of 'target audience' in the scoring rubric.

Discussion

This study was conducted to find out if there were differences in the scores given by two Mexican raters to the same papers written in English as a foreign language. Their views on what writing ability is and their views on the use of scoring rubrics to assess writing were also explored.

Significant differences were found among Raters' scores in spite of the similarity in their professional backgrounds. First, Rater A assigned higher scores than Rater B. Second, Rater A´s ratings seemed to be more varied and less consistent, while Rater B had less variance in his scores. These findings suggest that rubrics are not enough to produce homogeneous scores and that assessment judgment is influenced by distinct factors.

Results of this study comply with those found by Wiseman (2012) in which two distinct types of scoring, lenient and strict, depended on raters´ background and perceptions of text and task prompt. In this study, it is the researchers´ belief that rater score variation (one stricter than the other) depended on participants´ perceptions of the concept of writing and what is expected from students at different levels of proficiency. Therefore, it may be concluded from these results that raters´ perception of writing and student expectations are also part of scoring variation. The rater who associated writing ability with good use of grammar was more lenient than the rater who associated writing ability with discourse and sociolinguistic competence. Also, Rater A focused more on only one of the writing traits included in the rubric, while Rater B focused on a writing trait that was not included in the rubric. These results support the claim that the use of holistic or analytic scoring rubrics does not avoid the influence of a variety of factors in the raters' scoring behavior (Shi, 2001).

In this study, raters had similar background traits but differed in their perceptions of writing and written tasks. Results of this study suggest that background characteristics of raters may not influence score variability, but instead their rationale of writing and what to expect from writers. Therefore, these results can be differentiated from those found by Mendelsohn and Cumming (1987) and Santos (1988) in the personal and professional background of raters had influence on rating variability. Additionally, this study echoes the results from Barkaoui´s study (2010) in the sense that we believe that teaching experience was not a determining factor in rating variability. Instead other factors had more impact on variability than teaching experience. The researcher found that type of rubric used impacted variation while we consider raters´ personal perceptions to be determinant.

Finally, although only inter-rater variability was analyzed in this study, intra-rater reliability is an issue that should also be approached. As demonstrated in the study by Saxton, Belanger and Becker (2012), both types of rater analysis could be compared to explain the dynamics of both processes. Therefore, future research could focus on explaining the intra- and inter-rater assessment processes of specific writing tasks in an EFL context and compare them with others found in distinct language programs in the same context and seek to explain the variances found to suggest possible ways of obtaining fair and accurate ratings. Additionally, it could be worthwhile to focus on the

impact of different types of scoring rubrics and how these types of rubrics can result in more reliable ratings.

Conclusions and Implications

Results of this small-scale exploratory study allowed us to conclude that the use of scoring rubrics to assess EFL writing does not ensure homogeneity in an assessment process. Furthermore, similarity in raters' professional background, teaching experience, age and first language does not always imply similarity in their scoring behaviors. Other factors such as writing instruction and writing assessment training could possibly lessen the variability in scorings as found in studies by Saxton, Belanger and Becker (2012); Weigle, (1994) and Cushing (1994). These variables, however, were not within the scope of our study.

The variability found in the raters' scoring behavior has at least two implications for EFL writing teachers. First, the use of rubrics to assess the students' writing ability is useful for both students and teachers because it may enhance the quality of instruction and they may get students to think about the criteria on which their work will be judged. Scoring rubrics, however, are not a panacea. Ultimately, rating is complex and human behavior is multi-determined. The second implication is that although rubrics facilitate grading and communicating expectations to students, no scoring rubric fits all needs. Deciding on the best rubric to use for a specific writing task involves deep analysis of program goals and above all it involves training raters in its use.

Finally, this study involved two raters and therefore results cannot be generalized to other situations. The study, however, could shed some light on the issues involved in EFL writing assessment and the use of scoring rubrics. This insight can aid other instructors in comparing their specific context to that described in this study and seek for the best results in their everyday assessment practice.

References

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

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 6*, 152-163.

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

Cresswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage Publications.

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

Cushing, S. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.

Engelhard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5*(3), 171-191.

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

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.

Johnson, D., & VanBrackle, L. (2012). Linguistic discrimination in writing assessment: How raters react to African American "errors", ESL errors, and standard English errors on a state-mandated writing exam. *Assessing Writing, 17*(1), 35-54.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3-31.

Mayford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 518-575). Maple Grove, MI: JAM Press.

McDonough, J. & McDonough, S. (1997). *Research methods for English language teachers*. London, UK: Arnold.

McNamara, T. F. (1996). *Measuring second language performance*. New York, USA: Longman.

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

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

Pearson, P.C. (2004). *Controversies in second language writing: Dilemmas and decisions in research and instruction*. Michigan, USA: 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.

Santos, T. (1988). Professors´ reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly, 22*, 69-90.

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 nonnative 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*, 197-223.

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

Appendices (available online)

Appendix 1

Appendix 2