Jeong *Language Testing in Asia* (2015) 5:6
DOI 10.1186/s40468-015-0013-5

Language Testing in Asia
a SpringerOpen Journal

RESEARCH Open Access

Rubrics in the classroom: do teachers really follow them?

Heejeong Jeong

Correspondence:
jeongheejeong@gmail.com
College English Education Committee, Hanyang University , 222 Wangsimni-ro, Seongdong-gu Seoul 133-791, Republic of Korea

Abstract

Background: For language teachers, using rubrics has become the norm in assessing performance-based work. When using rubrics, one question stakeholders have is to what extent teachers are true to the rubrics. For classroom teachers, the correct use of rubrics is crucial. Rater training and rater calibration are not commonly offered to teacher-raters; therefore, the accurate use of rubrics is required in assessing student performance.

Methods: This study investigates the impact of rubric use in assessing short EFL descriptive writing by asking teacher-Q4raters to rate essays, both with and without a rubric.

Results: The results show that teachers focused more on errors (e.g., grammar and mechanics) when rating without a rubric, but valued comprehension issues (e.g., main idea, author's voice) when rating with a rubric. Essay scores also increased when teachers assessed with a rubric. Follow-up teacher interviews confirmed that rating changes occurred due to both the assessment criteria in the rubric and the lenient nature of the scale descriptors.

Conclusions: For performance-based assessment, rubrics are a central tool that adds reliability, validity, and transparency to assessments. This study shows that experienced teachers-raters were impacted by the content and nature of the rubric's scale, and thus made an effort to follow it.

Keywords: Teacher-rater; Rubric effect; Rubric literacy; Rubric fit

Background

Rubrics are used daily in the classroom, from simple scoring rubrics such as checklists, to more complex and detailed rubrics for final course projects or end-of-semester performances. Rubrics are especially valued in the language classroom, because they contribute to student learning and bring transparency to the assessment process (Wolf & Stevens, 2007). When performance based assessment began to receive attention, many rubric studies focused on covering the benefits of using rubrics (Andrade 2000; Jonsson & Svingby, 2007; Silvestri & Oescher, 2006; Spandel, 2006); how it promotes students' learning and makes performance based assessment more accurate and reliable. However, with the increase of using rubrics, the limitations of rubrics (Popham 1997; Kohn, 2006; Andrade 2001,) became the interest of many researchers. When students are given rubrics, they might question whether the teacher-rater is assessing them based on the rubric, or whether they are being graded on the teacher-rater's overall impression.

The rating process is still vague and teachers have been criticized for basing assessment on their overall impression (Lumley, 2002) Even though previous studies have found using a rubric adds more confidence in a teacher's rating (Silvestri & Oescher, 2006), rubrics themselves have been criticized for inconsistent criteria descriptors and vague language (Tierney & Simon, 2004). Popham (2003) warns, not all rubrics are well written and developing reliable and valid rubrics requires expert knowledge (Callison, 2000).

In the rating process, various factors come into play: rater characteristics towards severity or leniency (Schaefer, 2008; Shi 2001), rater training experience (Huot, 1990; Weigle, 1998, 2002), rater's language background (Kondo-Brown, 2002; Lumley & McNamara, 1995), and task variability (O'Loughlin and Wigglesworth 2007) are factors that have been researched over the years in performance assessment. Past research on rubric studies has focused on investigating changes in rater reliability (Lumley & McNamara 1995; McNamara, 1996; Weigle, 1998). Studies that show increases in rater reliability by using rubrics may, in part, be due to the training raters go through than by merely using rubrics (Knoch et al., 2007). Of course, learning how to use rubrics is part of rater training; however, we cannot assume that all raters, and especially teachers who are also raters, all receive professional rater training (Knoch et al., 2007). Evidence of rubric effectiveness can be shown through differences in inter-rater and intra-rater reliability (Jonsson & Svingby, 2007). However, measuring inter-rater reliability is only possible when there are two teachers, and for classroom teachers, having two teacher-raters is uncommon.

This study examines how teacher-raters are affected by rubrics by looking at how rating patterns change, with and without using a rubric. The purpose of the study is to research how teacher-raters respond, interpret and use rubrics in assessing EFL students' short descriptive writing, and to what degree it impacts students' grades. Through this study, I hope to determine whether teacher-raters work toward being true to a given rubric, or whether their ratings are still based on overall impressions.

Research on rubric effect

Rubrics are used to increase transparency in rater judgments and to decrease subjectivity (Silvestri & Oescher, 2006). Empirical evidence in rubric studies show that using rubrics can make assessments more reliable (Jonsson & Svingby, 2007; Penny, Johnson, & Gordon, 2000; Silvestri & Oescher, 2006; Wolf & Stevens, 2007). Most rubric-related literature has covered the advantages and benefits of rubrics (Jonsson & Svingby, 2007; Silvestri & Oescher, 2006; Spandel, 2006), but rubrics have also been criticized, in that they may limit raters' judgments and even limit the learning process (Turley & Gallagher, 2008; Wolf & Stevens, 2007). Kohn (2006) states teachers use rubrics to standardize their judgments regarding students work. He (2007) notes, "standardizing assessment for learners may compromise learning (p.13)." Popham (1997) also expressed concerns about rubrics in a similar vein stating that rubrics are used to assess students' "constructed responses (p.72)". Wolf and Stevens (2007) p.13 warns students might think "If it is not on the rubric, it must not be important or possible". The language used in rubrics has been criticized as being vague and unclear (Turley & Gallagher, 2008), which results in different interpretations by different raters (Knoch, 2009; Weigle, 2002).

Although most studies concerning rubrics agree that using rubrics will improve rater reliability (Jonsson & Svingby, 2007; Silvestri & Oescher, 2006; Penny, Johnson, & Gordon

2000), there is also a concern that merely having a rubric does not automatically add reliability and validity to an assessment (Tomkins, 2003). More researchers are stressing the importance of education and training in how to use rubrics (Turley & Gallagher, 2008; Wilson, 2007).

Studies on rubric use or its impacts on ratings are steadily increasing, but are still at a primitive stage. Previous rubric studies concentrated on the development and benefits of rubrics (Jonsson & Svingby, 2007); therefore, studies describing rubric effects on teacher-raters were difficult to find (Rezaei and Lovorn, 2010), and few studies covered the actual effects of their use. Empirical studies concerning rubrics have largely focused on reliability issues (Jonsson & Svingby, 2007), such that the main focus in these studies has been on rater agreement rather than on the rubric itself. Research conducted on how rubrics influence raters' judgments is limited. Despite many studies done on raters (Brown et al., 2004; Stemler, 2004), the process that raters go through in making final decisions using rubrics is still unclear (Huot, 1990; Lumley 2002; Lim 2011) and there are not many studies that look into the impacts of rubrics on raters.

A study that directly examined the effect of using rubrics compared to not using them when assessing writing was done by Rezaei and Lovorn (2010) with L1 graduate-level social science writing samples. The raters in this study first rated two student essays without a rubric, and later with one. The purpose of the study was to investigate the extent to which incorporating a rubric helps prevent raters from paying too much attention to the mechanical aspects of writing over content. The rubric used in the study put a strong focus on organization and content (75%) compared to mechanics (10%). Raters in the study were asked to rate an essay that was labeled as "correct" and another that was categorized as the "wrong" essay. The "correct" essay fully answered the prompt, but had frequent grammatical and mechanical errors, while the "wrong" essay did not fully answer the prompt, but was well polished and edited. The results show that, regardless of the "correct" or "wrong" essay, raters gave significantly lower scores when ratings were done with a rubric, and the range and variance of the assigned scores increased significantly, as well. The findings show that the inexperienced raters in the study were strongly influenced by the mechanical and grammatical aspects of students' writing, and this focus did not change by rating with a rubric. It is difficult to say whether the raters in this study closely followed the rubric in rating the essays. Grammar and mechanical factors still strongly impacted the raters' judgment, regardless of the stated criteria in the rubric.

To investigate the impact of rubrics on teacher-raters, this paper will focus on the following areas:

1. What do teacher-raters value in assessing short EFL descriptive writing?
2. Do ratings change when using a rubric?
3. What are the reasons behind the changes or the decision not to change?

Methods

Data for this study came from teacher-raters' essay ratings, rating justifications and interviews. Essay ratings were checked to identify rating changes, with and without using a rubric (Figure 1). Rating justifications were analyzed to have an idea of the assessment constructs of the teacher-raters. Teacher-rater interviews were conducted to explain the rating changes. This study used a mixed-methods approach, using both

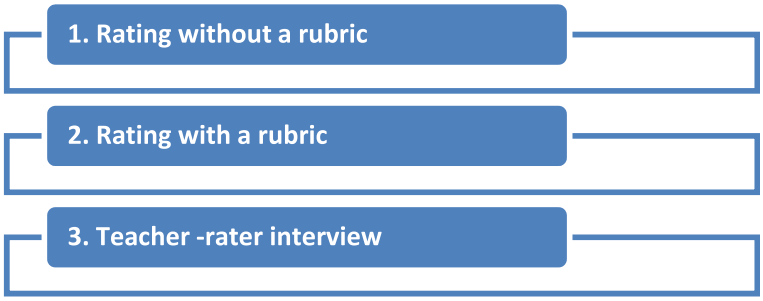

Figure 1 Data collection process.

quantitative and qualitative methods to strengthen the methodology for this study. The purpose of mixing was for complementarity reasons. The complementary nature of the design strengthened the instruments used for this study. The qualitative findings from the rater interviews complemented the quantitative findings of the essay ratings.

First, each teacher-rater was given 20 student essays via email and was asked to rate short descriptive essays written by Korean EFL writers (Additional file 1) without a rubric. Teacher-raters were asked to grade the essays on a 6-point letter scale: *A+, A, B+, B, C+, C* and write a justification of the rating next to their letter grade explaining the reasons for the rating. Teacher-raters' essay justifications were later analyzed to determine the reasons associated with the score when rating without a rubric and to identify valued assessment criteria. The ratings and justifications were sent to the researcher prior to the interview.

During the interview (Additional file 1), teacher-raters were given the Rubric (Additional file 1), and were asked to do a second rating for five selected essays from the original 20 essays. Two essays were rated by all raters and the three were randomly chosen. The purpose of the second rating was to check whether rating changes occurred when raters used the Rubric. When all ratings were completed, an in-depth interview was conducted to compare ratings, with and without a rubric, and the reasons behind the changes or the decision not to change the rating. The interviews were audio-recorded and transcribed by the researcher.

Teacher-raters

The teacher-raters for this study were native EFL instructors from a large private Korean university. All were experienced teacher-raters who had a wide range of teaching experience from 5 ~ 12 years in higher education. The participants' rating experience was mostly within the university context. The courses they taught were graded on students' performance-based activities; moreover, they had experience in rating the English placement test that was conducted every year for freshman students. A few had experience taking part as a judge for an essay contest and presentation contest held on campus.

The rubric

The Rubric (Additonal file 1), which is the given rubric designed by the researcher, was based on previous writing rubrics used by forty teachers from the target language institute. Prior to developing the Rubric, rubrics that were currently used to assess EFL

descriptive writing were collected and analyzed. Frequently used assessment criteria were compiled and incorporated into the Rubric. Writing criteria rubrics from large-scale English tests (e.g., IELTS, TOEFL) were also analyzed. The overall main assessment criteria came from teacher rubrics and well-known essay rubrics but one criterion (i.e., author's voice) was intentionally added for the purpose of the study. The Rubric was two pages long and used a scale ranging from *A+, A, B+, B, C+, C*. The Rubric consisted of seven scoring criteria: main idea, support, flow/cohesion, author's voice, sentence type and variety, vocabulary and word choice, and grammar and mechanics. The Rubric highly emphasized comprehension (e.g., main idea, support, author's voice) issues rather than sentence- level (e.g., sentence structure, vocabulary) or editing concerns (e.g., grammar, mechanics). The left-hand side of the Rubric contained the assessment criteria and the right-hand side included descriptors that detailed each criterion for each grade level. The descriptors in the Rubric were consistent across levels; thus, descriptions that appeared in one level (e.g., strong author's voice) were present in all levels (e.g., lack of author's voice). The first criterion, "main idea" discussed the originality, clarity and intelligibility of the idea; the criterion "support" covered the degree of elaboration and development of supporting details. The third criterion, "flow/cohesion," involved the logic and organization of the paragraphs, while the next criterion, "author's voice," discussed the strength and clarity of the author's voice. The fifth criterion, "sentence type and variety," explained variation in sentence length and structure; this criterion also assessed grammatical points such as sentence fragments and run-on sentences. The sixth criterion was "vocabulary and word choice," which concerned the effective use of words and the use of descriptive language. The final assessment category was "grammar and mechanics," which described general grammatical errors and correct usage of capitalization, punctuation and spelling. Compared to the rubric that was formerly used to assess similar writing tasks, the Rubric was intentionally designed to be more complex and covered a foreign assessment criterion (e.g., author's voice) to probe more discussion from the teacher-raters. I also purposely put a strong emphasis on criteria related to essay comprehension rather than sentence structure and grammar to see whether there would be any changes when teacher-raters rated essays with the Rubric.

Data analysis

Essay ratings were submitted as a letter grade and were converted to numerical equivalents ranging from 1 to 6 for the statistical analysis. The essays that were rated twice were compared for each rater and for each essay using paired *t*-tests using SPSS V.21, and the maximum and minimum ranges of scores were noted. Teacher-rater essay justifications from rating without a rubric were analyzed in order to identify the teacher-raters' assessment criteria of assessing this task. Essay rating justifications were summarized and condensed by using descriptive coding. Through this analysis, nine assessment criteria were identified: comprehension (COM), paragraph structure (PS), sentence structure (SS), vocabulary (VOC), grammar (GM), mechanics (MC), length (LNT), task completion (TC) and self-correction (SC). Frequencies of the rating justifications for each category were counted across teacher-raters to detect the key criteria used for assessing EFL writing. Teacher-rater interviews were digitally recorded and transcribed for thematic analysis. The interview analysis focused on reasons for making rating changes, meaning of the score and the rubric fit.

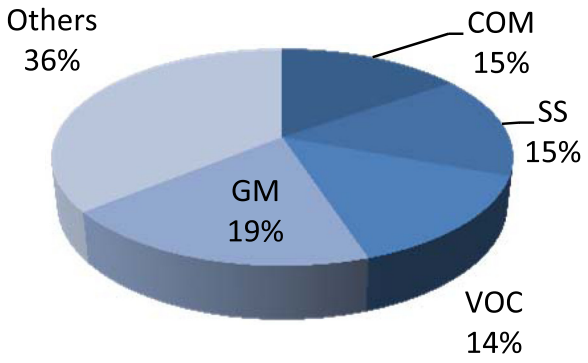

Figure 2 Without rubric: key assessment criteria.

Results

What do teacher-raters value in assessing short EFL descriptive writing?

The first step of the study was to find the key assessment criteria used by teacher-raters in assessing EFL descriptive writing. This step was needed in order to compare what teacher-raters valued when rating without a rubric, compared to rating with a rubric. The above figure (Figure 2) shows the percentage of commented criteria from the essay justifications when rated without a rubric. The number of times each criterion was mentioned for the 20 essays was calculated for each teacher-rater. The findings show that teacher-raters valued GM (19%, n = 70) as the most important criterion for assessing short EFL descriptive writing. Following this criterion was COM (15%, n = 56), SS (15%, n = 56) and VOC (14%, n = 52).

Do ratings change when using a rubric?

For the second rating, five essays from the original twenty were rated by teacher-raters according to the Rubric. Out of the 25 ratings, 72% (n = 18) of the essays were given a different score (Figure 3). Only 28% (n = 7) showed no changes in the ratings.

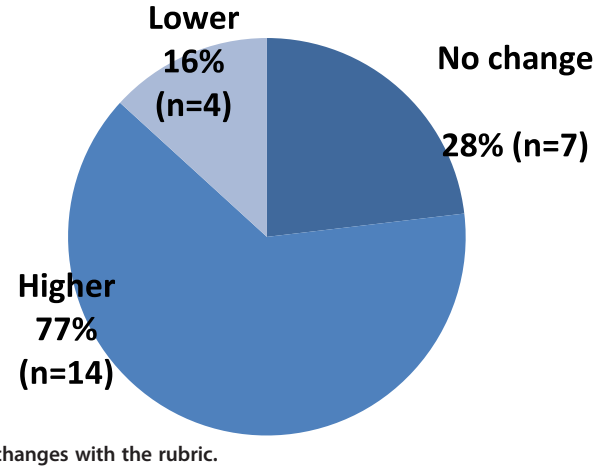

Figure 3 Rating changes with the rubric.

Table 1 With and without the rubric: rating scores

Teacher-rater	Without rubric (n = 5)			With rubric (n = 5)			*t*	*p*
	Mean	SD	Min/Max	Mean	SD	Min/Max		
Eunice	2	.707	1/3	2.4	1.140	1/4	−.590	.587
Susan	3.2	1.304	2/5	3.8	1.304	2/5	−2.449	.070
Ben	2.8	.447	2/3	3	.707	2/4	−1.0	.374
Logan	2.6	.894	2/4	3.6	1.817	1/6	−1.826	.142
Matt	2.8	.447	2/3	3	.707	2/4	−.535	.621

Rating changes by raters were usually within a score point and the average essay score was higher for all teacher-raters (Table 1). The score range and variance of the scores increased after rating with the Rubric but none of the teacher-raters showed a significant difference in scores when rating with or without the Rubric. The rating done with the Rubric was higher for all raters but none showed a statistically significant difference; Eunice, $t(4) = -.590$, $p = .587$, Susan, $t(4) = -2.449$, $p = .070$, Ben, $t(4) = -.534$, $p = .621$, Logan, $t(4) = -1.826$, $p = 1.42$, Matt, $t(4) = -.535$, $p = .621$. Among the teacher-raters, Logan showed the biggest increase for the mean score and the widest score range after rating with the Rubric. Susan showed no changes in the score range, but the other three teacher-raters' maximum point increased by a point.

Two essays (#2, #13) that were rated twice by all raters showed a significant increase in scores (Table 2). When essay #2 was rated with the Rubric it received an overall score of 3.4 which showed a significant increase, $t(4) = -3.162$, $p = 0.034$. This was the same for essay #13, $t(4) = -6.0$, $p = 0.004$.

What are the reasons behind the rating changes or the decision not to change anything?

Assessment criteria

During the interview, each teacher-rater was asked to describe the reasons behind the rating changes or the reason not to change (Table 3). Eunice, commented that the Rubric deals with fresh ideas and the author's voice, which she did not consider much when rating without a rubric. When asked to provide reasons for any changes in her ratings, Eunice stated that she did not initially consider the strength of ideas as an important criterion in her initial rating for short EFL descriptive essays. She thought the reason may have been due to different amounts of attention given to specific grammar points when she described a specific essay to which she had made rating changes. Eunice stated, "[In my first rating], I was paying a lot of attention to the run-on sentences. With this [the Rubric], I was more interested in support. If the details were appropriately developed, even

Table 2 With and without the rubric: rating scores by essay

Teacher-rater	Without rubric (n = 5)			With rubric (n = 5)			*t*	*p*
	Mean	SD	Min/Max	Mean	SD	Min/Max		
Essay 2	2.4	.547	2/3	3.4	.547	3/4	−3.162	.034*
Essay 13	2.2	.447	2/3	3.4	.547	3/4	−6.0	.004*

*p<.05

Table 3 Changes in assessment focus when rating without or with a rubric

	Without rubric	With rubric
Eunice	discrete grammar points, parts of speech, capitalization	idea, author's voice
Susan	idea	idea
Ben	sentence structure, form	author's voice
Logan	tense usage	organization, flow, cohesion, transitions
Matt	errors (verb tense agreement) articles or prepositions, capitalization, and punctuation	less on errors

though the sentences were run-on, [it was okay]. I think they could have been fixed pretty easily." Eunice thought her ratings increased because she focused on the idea or support of the idea rather than on other sentence-level concerns, such as sentence variety. "I think [I am more generous] because of the ideas and details. I was trying to see the author's voice, less on parts of speech or capitalization," she explained.

Ben felt he was influenced by the Rubric, when explaining the rating changes of the same essay. For an essay which he originally gave a *C*+ (2 points) and in the second rating changed to a *B*+ (4 points) he said, he focused on form rather than content when rating without a rubric. Ben states, "I probably put more emphasis on the fact it wasn't a paragraph [when I rated without the Rubric]. [I found] the sentence structure very simple and there are errors [so] I gave that a *C*+. [When I rated it for the second time] I felt it [the Rubric] influenced me. I was very close giving that a *C*+, but the writer had an author's voice. There is some evidence of author's voice there. She talked about the college entrance exam but it was about an argument with her mother, so she is trying to get to something that is more personal. It kind of felt there was a little more there. So that is why I gave it a *B*+." For essays, he did not make changes in the ratings; he commented that the focus of the assessment criteria was different. Ben felt that the Rubric emphasized completely different things from what he had initially looked for; therefore, the meaning of a "B" grade, in his mind, was different even though he had assigned the same scores when rating, with or without a rubric.

In the interview, Matt commented, "[When I rated without a rubric], I really focused on types of errors, actually pinpointed to detailed error types, for example, incorrect verbs." Matt stated that he focused on common EFL errors concerning articles or prepositions, capitalization, and punctuation errors for the first rating. Yet, he noticed in the Rubric that grammar errors and mechanics were combined together as a single criterion, so he could not strongly focus on each category, as compared to before.

Logan thought that the rating changes occurred due to the focus on the organization, flow and cohesion criteria specified in the Rubric. He said, "According to this [the Rubric], in terms of ability to organize, express herself, she was very good, very good transitions. I think that is why [I changed my rating]." For an essay, which he changed from a *C*+ to a *B*+, he commented, "I gave her a *C*+ [in my first rating]. I think I got the general gist, but I have to go back and think about it. From this rubric, [the essay] shows an attempt to organize a paragraph,

and there were some transitions. So according to these indicators that's why I gave it a *B*+."

Unlike other teacher-raters, Susan did not think rating changes occurred due to the different assessment focus of the Rubric; in fact, Susan believed that the Rubric's criteria were similar to her own assessment focus, which was strength of the ideas.

Scale severity

For Susan, her ratings changed because of the difference of the scale severity. She stated that her reason for giving a more generous score was due to the lenient nature of the *C*-level description in the Rubric. She stated, "For the *C*-levels, I'm probably harsher than what you are looking for [in the Rubric]. Your *C* is my *C*+, your *C* is literally [main ideas that] do not exist. For my *C,* there still has to be something on the paper." She then added, "If I were to re-do the twenty again [following the Rubric], my *C*s will be *C* + s."

Other teacher-raters also talked about scale severity in the Rubric. Logan believed that some of the rating changes occurred because of the lenient nature of the Rubric; in specific, Eunice commented that the Rubric had higher expectations for the *A*+ level than her own. She stated that the Rubric was strict on the high end, but generous on the low end.

Discussion

The rating changes did not show a statistical significant difference by rater but was statistically significant for the two essays (#2, #13) that were rated twice by all five raters. This different result could be due to the content of student essays. Essays that were weak in grammar and structure but strong in ideas could have resulted into a score change when rating with the Rubric. All raters except Susan, confirmed in the interview their assessment focus shifted from grammar to content when rating with the Rubric. However the findings from the *t* tests should be interpreted with caution considering the small sample size of raters (n = 5) and lack of essays (n = 2) that were rated twice by all raters.

In terms of the meaning changes associated with the score, findings from this study imply that rating changes can be explained by two different rubric effects; assessment criteria and scale severity. For four teacher-raters, the first ratings primarily focused on errors concerning grammar and mechanics. This phenomenon is consistent with previous studies that conclude when raters rate without a rubric, they are highly influenced by grammar or editing mistakes (Read, Francis, & Robson 2005). The reasons as to why teacher-raters in this study focused on these areas could be due to the style and content of the writing. The writing samples were 1 ~ 3 paragraphs in length, and at times, were underdeveloped. As Eunice commented, for this type of writing task, she focused more on sentence-level errors rather than on the development of a paragraph or an idea. In other words, the teacher-raters could have paid more attention to errors because they did not have a rubric or because those criteria were a better fit for the given task.

Regardless of what influenced the assessment construct of the raters, they showed rating changes or evidenced different meanings associated with the

ratings after using a rubric. Teacher-raters were directly influenced by the assessment criteria included in the Rubric. All teacher-raters, except for Susan said that the Rubric focused more on the strength and development of ideas, compared to what they had initially expected. For Susan, her top assessment criterion was idea development; therefore, there was no difference between her understanding of this criterion from that of the Rubric. Teacher-raters changed their ratings after following the descriptors in the Rubric and found evidence to support their ratings. The criterion "author's voice" had an impact in changing the ratings for Eunice; moreover, Eunice, Logan, and Matt all reported putting less focus on grammar and mechanical errors, which resulted in an increase in their ratings.

An interesting finding with respect to why Susan changed her rating deals not with the difference in the assessment criteria, but with the generous nature of the rating scale. Susan felt that especially for the *C+/C* level descriptors, the Rubric was much more lenient, compared to her original expectations. Differences in the assessment criteria could be the biggest factor in rating changes for teacher-raters, but differences in the scale severity also had an influence.

Even though the Rubric could have strongly influenced the rating changes, this probably is not the sole factor. Rating changes could have also occurred due to the rating style or characteristics of the teacher-rater. Eunice noted that she is quite lenient whenever she is given a new rubric. She said that it takes time to understand and use a new rubric, and during the learning period, she tends to give higher scores. The level of confidence that teacher-raters have when rating without a rubric may also be a factor. Ben and Mark expressed extreme difficulties when rating without a rubric, and they questioned their judgments. They felt insecure about rating without a scale and believed they lacked consistency in their first ratings. Time can also be an important factor in rating changes, in that the teacher-raters could have given different ratings using the same rubric merely due to the difference in the timing between the ratings.

Nonetheless, the findings from this study indicate that raters are influenced by rubrics to some extent. This finding is different from Rezaei and Lovorn's (2010) study, which found that using rubrics did not take away raters' focus on trivial mechanics or superficial aspects of writing. The raters in Rezaei and Lovorn's study showed a wider variance and score decrease after rating with a rubric. This study, on the other hand, showed a score increase when essays were rated with a rubric. When rating with the Rubric, the maximum score for most teacher-raters did increase which reflects the generous characteristic of the Rubric. One main difference between Rezaei and Lovorn's study and the present one is the background of the raters and the characteristics of the writing. The raters in this study were experienced teacher-raters who have had many years of experience rating with a rubric. Their understanding of rubrics is likely to be higher than that of non-experienced raters. Although the teacher-raters in this study did not receive any formal rater training for this study, in the interview they did comment on receiving regular rating training once or twice a year from the program they worked at. Previous studies in rater training have found that rubric training helps raters focus on the content of the rubric and helps them discard their personal biases during the rating (Knoch et al., 2007). Experience in essay rating and using rubrics could be the reasons for the divergent results.

Following the criteria and scale in the rubric does not imply that the teacher-raters' assessment constructs changed or that they agreed the Rubric was the best tool for assessing the given task. In fact, Eunice and Logan expressed dissatisfaction in the Rubric and thought that it did not fit well for assessing short descriptive EFL writing. Specifically, Logan thought that the Rubric was too long, complex and difficult to use. In addition, Eunice believed that the Rubric should put more emphasis on vocabulary and sentence-level issues rather than on the strength of ideas. Ben and Matt did not like the fact that the Rubric covered grammar and mechanics under the same category; both wanted to split the categories and put a stronger emphasis on grammar. Despite their dissatisfaction with certain areas of the Rubric, the teacher-raters in the study still made an effort to follow it. For them, the Rubric overrode their personal assessment beliefs. In the interview, Susan commented, "I usually follow the rubric in the order it is given to me, because I think that is what I am asked to do. I do not put my two cents in. I try to stay true to the intent of the rubric, without any embellishment on my part." Susan also added that when giving grades she tells her students, "It wasn't me who gave you this grade, the rubric did".

Raters may be heavily influenced by their overall impression, but the experienced teacher-raters in this study tried diligently to follow the given the rubric. They were aware of their role as a rater, and even though they may not have accepted the descriptions and criteria in the Rubric, they tried to follow it. Personal intuition could have still played a role in their decision-making process; however, the teacher-raters worked toward making their decisions based on the criteria of the given rubric.

Rubric literacy

While other factors (e.g., time, assessment confidence, rating style) come into play in describing the rating changes, the impact of a rubric is definitely a crucial factor in guiding teacher-raters to make rating judgments. Teacher-raters in the study tried to be true to the Rubric and found evidence within the document to support their ratings. A rubric is a complex scoring guide consisting of criteria, descriptors and scales. To be used correctly, teacher-raters should be educated in how to use a rubric. Rater training should not only focus on rating practices, but prior to doing ratings, sufficient time should also be given in learning the rubric. Clear explanations should be given for the meaning of each criterion and examples of the descriptors. For example, the criterion "author's voice" was foreign to the teacher-raters in this study, and was puzzled in how to interpret this criterion. Susan felt that if a person writes something, the essay automatically contains the voice of the author; on the other hand, Eunice and Ben believed it relates more to how creatively the writer expresses his/her thoughts and ideas in the essay.

Susan stated that there is a "learning curve" in adapting to every new rubric. Teacher-raters should form groups and discuss the language and criteria in the rubrics; in addition, they should clarify any questions with respect to meaning. Rubric literacy, which is part of a teacher's assessment literacy, is a crucial part to being a reliable teacher-rater. What is more important than developing and using rubrics is how to understand and interpret them. In addition to developing one's knowledge in how to use rubrics, teacher-raters should be trained in how to select an appropriate rubric for a given task. Rubric literacy should cover overall rubric fit in terms of how to select, analyze, develop, and use rubrics.

To help teachers to have a better understanding in developing and using rubrics language program directors should encourage teachers to discuss their expectations of each grade level or task prior to assessing student work. Teachers can get together and develop or adapt a rubric for the same task. Teachers can develop their own rubric and later compare and contrast the criteria, scales, and rubric style. Based on the similarities and differences of individual rubrics, a standard rubric that can be used across all teachers can be produced. This method can give an opportunity to discuss and visualize teachers' assessment constructs and develop common criteria for a grade or score. After a rubric is developed together, teachers could have a workshop practicing using the rubric. Rubric literacy is not only about possessing the knowledge to develop good rubrics but also knowing how to use them correctly (Andrade, 2005). Rubric practice sessions could be done among teachers and through the practice session a document that describes each criterion can be produced (Andrade, 2005).

For large-scale assessments, there are multiple types of evidence to demonstrate the reliability and validity of a rating. Validity reports can be produced and different psychometric analyses can be done to show that the given score is reliable and valid. In a classroom setting, where the teacher is the only rater, however, these methods cannot be incorporated. Data that can back up a teacher-raters' judgment can come from rubrics. A rubric justifies the ratings of teacher-raters and adds objectivity to their judgments. If an appropriate rubric is selected and teacher-raters correctly use them, the rating based on the rubric will be accepted as a valid and reliable score.

Limitations and directions for future research

One of the limitations of this study is that students had two choices in writing the descriptive essay. The reason for giving a choice was to lower the anxiety level of the students and make it more accessible, yet, there is a concern that allowing task choices can bring difficulties in analyzing test results which can be a threat to reliability and validity. However, the purpose of this study was not focused on the use of essay scores but on how raters responded to rubric use. For future studies it would be better to give a single prompt to resolve measurement difficulties.

Another limitation was the few number of essays that were rated for the study. Only twenty essays were given to raters for the first rating and five were given for the second rating. Among the five, only two were rated twice by all raters. This study used both quantitative and qualitative methods but attention was given more towards to the qualitative data. In order to secure sufficient time during the teacher-rater interviews and write up of the rating justifications, a limited number of essays were used. In the future, I suggest studies to increase the number of essays and raters to strengthen the research findings.

Conclusion

This study investigated the impact of using rubrics by comparing essay ratings when rated with or without a rubric. The findings show despite teacher-raters' different assessment constructs, when they were given a standard rubric all

teachers made an effort to follow it. The teachers in this study had extensive experience in using rubrics. The results from this study show compared to novice raters (Rezaei and Lovron, 2010), experienced raters knew how to use a given rubric. The criteria stated in the rubric overrode their personal assessment constructs. Teacher-raters in the study were clear of what was expected to them as a rater and knew how to use a rubric and rated accordingly to the stated criteria. Thus, experienced teacher-raters did not base their ratings on their overall impressions but followed the given rubric. This finding shows the importance of rating experience for teacher-raters and training in how to use rubrics. If regular training is not available for classroom teachers, at least a document that explains how to use a given rubric should be presented.

The increase of performance-based assessment is a universal phenomenon for language classes around the world. Technological improvements have made it easier to document students' productive skills. Moreover, technology-based assessment has developed over the years, such as automated scoring for writing or speaking tests. The acsmdvancement of computer-based assessment has definitely brought about changes in assessments; nevertheless in the classroom context, the role of the teacher-rater is vital. For performance-based assessment, rubrics are the central tool to add reliability, validity, and transparency in classroom assessment. Despite the conflicts with their personal assessment constructs, the teacher-raters in this study worked diligently in terms of following the Rubric.

Additional file

Additional file 1: Appendix A. Rating Instructions. **Appendix B**. Teacher-rater Interview Questions. **Appendix C**. The Rubric.

Competing interests
The authors declare that they have no competing interests.

Received: 14 October 2014 Accepted: 27 January 2015
Published online: 15 April 2015

References

Andrade, HG. (2000). Using rubrics to promote thinking and learning. *Educ Leadersh, 57*(5), 13–18.
Andrade, H. G. (2001, April 18). The effects of instructional rubrics on learning to write. *Current Issues in Education* [Online], *4*(4). Available http://cie.ed.asu.edu/volume4/number4
Andrade, HG. (2005). Teaching with rubrics: The good, the bad, and the ugly. *Coll Teach, 53*(1), 27–30.
Brown, GTL, Glasswell, K, & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing system. *Assess Writ, 9*, 105–121.
Callison, D. (2000). Rubrics. *School Library Media Activities Monthly, 17*(2), 34,36,42.
Huot, BA. (1990). Reliability, validity and holistic scoring: What we know and what we need to know. *Coll Compos Commun, 41*, 201–213.
Jonsson, A, & Svingby, G. (2007). The use of scoring rubric: Reliability, validity and educational consequences. *Educational Research Review, 2*, 130–144.
Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Lang Test, 26*(20), 275–304.
Knoch, U, Read, J, & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assess Writ, 12*(1), 26–43.
Kohn, A. (2006). The trouble with rubrics. *Engl J, 95*(4), 12–15.
Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Lang Test, 19*(1), 3–31.
Lim, GS. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Lang Test, 28*(4), 543–560.
Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Lang Test, 19*(3), 246–276.
Lumley, T, & McNamara, TF. (1995). Rater characteristics and rater bias: Implications for training. *Lang Test, 12*(1), 54–71.

McNamara, TF. (1996). *Measuring second language performance*. London: Longman.
O'Loughlin, K, & Wigglesworth, G. (2007). Investigating task design in academic writing prompts. In L Taylor & P Falvey (Eds.), *IELTS collected papers. Research in speaking and writing performance* (pp. 379–421). Cambridge: Cambridge University Press.
Penny, J, Johnson, RL, & Gordon, B. (2000). Using rating augmentation to expand the scale of an analytic rubric. *Journal of Experimental Education, 68*(3), 269–287.
Popham, WJ. (1997). What's wrong-and what's right-with rubrics. *Educ Leadersh, 55*, 72–75.
Popham, WJ. (2003). *Test better, teach better: The instructional role of assessment*. Alexandria: Association for Supervision and Curriculum Development.
Read, B, Francis, B, & Robson, J. (2005). Gender, bias, assessment and feedback: Analyzing the written assessment of undergraduate history essays. *Assessment and Evaluation in Higher Education, 30*(3), 241–260.
Rezaei, A, & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assess Writ, 15*, 18–39.
Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Lang Test, 25*(4), 465–493.
Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Lang Test, 18*(3), 303–325.
Silvestri, L, & Oescher, J. (2006). Using rubrics to increase the reliability of assessment in health classes. *International Electronic Journal of Health Education, 9*, 25–30.
Spandel, V. (2006). In defense of rubrics. *Engl J, 96*(1), 19–22.
Stemler, SE. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4), 1–19.
Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels. *Practical Assessment Research and Evaluation, 9*(2), August 12, 2013.
Tomkins, M. (2003). Trouble comes in threes. *Times Educational Supplement, 4547*, 23.
Turley, ED, & Gallagher, CG. (2008). On the uses of rubrics: Reframing the great rubric debate. *Engl J, 79*(4), 87–92.
Weigle, SC. (1998). Using FACETS to model rater training effects. *Lang Test, 15*, 263–287.
Weigle, SC. (2002). *Assessing writing*. Cambridge: Cambridge University Press.
Wilson, M. (2007). Why I won't be using rubrics to respond to students' writing. *Engl J, 96*(4), 62–66.
Wolf, K, & Stevens, E. (2007). The role of rubrics in advancing and assessing student learning. *The Journal of Effective Teaching, 7*(1_2), 3–14.