Conventional and Rater-Friendly Rubrics for L2 First Year Composition

Kevin Hirschi and Chelsea Moreno

Northern Arizona University

Abstract

Rubrics have long been used to evaluate L2 student writing. However, little research has been done on rubrics for L2 student essays at the undergraduate first-year composition level (Becker, 2010; East & Young, 2007). Because of the unique differences found in L2 student writing, a new rubric, referred to as the *rater-friendly rubric*, was created in a freshman composition course for upper-intermediate L2 students at a Northern American university. This new rubric focused on the accomplishment of tasks set out by the essay prompt, to allow teachers to rate as the anticipated task was accomplished in the student essay. This quantitative and qualitative ex-post facto study investigated the differences in rater reliability and rater confidence when comparing the former rubric and the new rater-friendly rubric. 13 raters evaluated eight essays and found that the new rubric to be more reliable than the former rubric through qualitative and quantitative analyses. Implications for teachers and student writing evaluators are given.

Conventional and Rater-friendly Rubrics for L2 First Year Composition

Background

L2 student needs and their essays are often very different than L1 students. When comparing the two, it may seem to be more difficult to reliably score L2 essays for a variety of reasons. Beyond simple errors that make meaning of essays difficult to interpret, corpus investigations into the syntactic and lexical differences have found that L2 writing at the university level tends to be less complex (Ferris, 1994; Lu, 2011). Furthermore, the approach, process, motivation, and discourse results can vary widely from L1 norms, resulting in lower scores (Atkinson & Ramanathan, 1995). These issues outline the need for a different approach to L2 writing evaluation at the first year composition level found in many Intensive English Programs (IEPs) (Becker, 2010).

Writing rubrics often take a conventional, analytic form in which they are divided into bands such as *Focus/Organization, Language, Elaboration,* and *Mechanics* (Miller, Linn, & Gronlund, 2013). A rubric following this method has been shown to result in low reliability ratings amongst instructors at the same institution as the current study with similar students (Yol, 2015*).* According to Miller et al. (2013), low reliability and construct validity in essay scores is due to a lack of construct definition and the absence of a quality rubric.

This was the impetus for the creation of new rubrics that assess the specific guidelines of the essay prompt and the focus of classroom instruction in the same order as those elements appear in the student's essay. These *rater-friendly* rubrics allow the rater to assess each requirement independently as well as the *introduction*, *thesis statement, conclusion, formatting, grammar and punctuation,* and *word choice,* elements of focus in the L2 writing classroom. This

study investigated the pilot results of using the new rater-friendly rubric by looking at differences in rating results between the conventional rubric and the rater-friendly rubric. It also set out to assess rater confidence in the scores they assign using both rubrics.

Research Questions

1) To what extent are *conventional* rubrics and *rater-friendly* rubrics reliable?
2) To what extent are raters confident about the scores they assign using *rater-friendly* rubrics or *conventional* rubrics?
3) Do overall grades vary between rubric types?

Methods

Eight final essays were collected from two intact groups of IEP students in the bridge first-year composition course. Students were placed in this level by a placement test. The two groups consisted largely of science, technology, engineering, and math majors. Essays were chosen from each group to represent a variety of performance levels. Raters were recruited from graduate students of English Rhetoric and Teaching English as a Second Language MA programs as well as Applied Linguistics doctoral students at the same institution. All raters were either familiar with teaching at the IEP or teaching the first-year composition course to L1 or L2 students. However, their amount of training in teaching, writing evaluation, and assessment with both L1 and L2 varied. All raters were native speakers of English. Quantitative and qualitative data was collected from the 13 raters evaluating 8 essays using electronic data collection methods. Each rater was given a randomly assigned order of the essays to evaluate as well as rubric type-to-essay combination.

The conventional rubric includes five bands of evaluation. One band was omitted as it included peer-review, which could not be evaluated through this study. The rater-friendly rubric

had of 12 evaluation bands (see Appendix). One band was omitted as it required knowledge of the peer review results and therefore was not feasible to include in this study. Essays were de-identified, and distributed to raters. Raters had two weeks to complete the rating process. All results were collected using an online form software, and were analyzed using Excel and IBM's SPSS.

Results

To answer Research Question 1, results for total scores was collected and analyzed using descriptive statistics. See Table 1 for results.

Table 1

Descriptive Statistics for Essays 1-8 Evaluated by Conventional and Rater-friendly Rubrics

	Conventional Rubric			Rater-Friendly Rubric		
	n	M	SD	n	M	SD
Essay 1	9	84.56	19.30	3	110.00	10.82
Essay 2	4	119.25	3.80	7	116.57	5.97
Essay 3	9	112.89	8.29	4	122.00	8.98
Essay 4	5	108.40	14.63	6	110.67	7.94
Essay 5	8	98.75	16.59	5	119.80	6.50
Essay 6	4	110.63	22.47	7	119.86	10.87
Essay 7	8	101.50	20.49	4	119.00	7.48
Essay 8	3	108.00	13.23	8	113.25	10.04
Average	6.25	105.50	-	5.5	116.40	-

To calculate reliability, two statistical procedures were carried out. Inter-rater agreement was calculated for each rubric band, then averaged for each essay and again averaged for each rubric type. The results showed an average of .86 rater agreement for the conventional rubric and .95 for the rater-friendly rubric. See Table 2.

Table 2

Inter-rater agreement for each essay by rubric type

	Conventional Rubric Inter-Rater Agreement	Rater-Friendly Rubric Inter-Rater Agreement
Essay 1	0.81	0.93
Essay 2	0.86	0.95
Essay 3	0.82	0.98
Essay 4	0.87	0.94
Essay 5	0.89	0.96
Essay 6	0.86	0.96
Essay 7	0.88	0.95
Essay 8	0.89	0.91
Average	0.86	0.95

A second measure of agreement was conducted to further illuminate rating differences between the two rubrics. Intraclass Correlations were calculated using SPSS software's two-way random analysis of intraclass correlation (ICC) for scale data ICC(2, 1). The ICC coefficient for the conventional rubric was calculated at .59 and the ICC coefficient for the rater-friendly rubric was calculated at .72, as shown in Table 3.

Table 3

Intra-Class Correlations for Essays 1-8 for Conventional and Rater-friendly Rubrics

	ICC	95% CI	$p <$
Conventional Rubric	.59	[.38, .75]	.00
Rater-Friendly Rubric	.72	[.58, .84]	.00

For Research Question 2, descriptive statistics were conducted on Likert scale results for all ratings using each rubric. In this scale, 4 indicates very confident and 1 indicates not confident. Table 4 indicates little difference between the two rubrics.

Table 4

Confidence Ratings of Conventional and Rater-friendly Rubrics

	Conventional Rubric			Rater-Friendly Rubric		
	n	M	SD	n	M	SD
Confidence	230	3.45	.63	480	3.51	.72

To answer Research Question 3, descriptive statistics were run. A paired-samples t-was used to determine if there is a statistically significant difference between the rubrics. Effect size was calculated to show score differences between the rubrics. Mean scores are summarized in Table 5. T-test results indicate a significant difference between the mean scores ($t_{critical}$ = +/-2.36, $t_{observed}$ = -3.19, p <.01). The effect size for the analysis (d = -1.34) was found to meet Cohen's (1988) convention for a medium effect (d = -.56).

Table 5

Mean Score Comparisons on Conventional and Rater-friendly Rubrics

	Conventional Rubric Mean	Rater-Friendly Rubric Mean	Mean Change in score
Essay 1 total score	84.56	110.00	25.44
Essay 2 total score	119.25	116.57	-2.68
Essay 3 total score	112.89	122.00	9.11
Essay 4 total score	108.40	110.67	2.27
Essay 5 total score	98.75	119.80	21.05
Essay 6 total score	110.63	119.86	9.23
Essay 7 total score	101.50	119.00	17.5
Essay 8 total score	108.00	113.25	5.25
Average total score	105.50	116.40	10.9

The validity of the rater-friendly rubric was confirmed by qualitative results. The raters indicated that the rater-friendly rubric more closely represented the actual quality of students' work than the conventional rubric. Raters also indicated that the conventional rubric assigned

scores that were too low, which may have been due to the increased weighting of language use on this rubric. However, some raters claimed that the rater-friendly rubric assigned scores that were too high because language use was weighted lower than content-related rows. Despite this, more raters claimed that the rater-friendly rubric more accurately represented the quality of writing than the conventional rubric.

When asked which rows were easier to score than others, raters indicated that rows that were specific and easily quantifiable were easier than rows based on more subjective judgment. This may have influenced raters' higher confidence levels regarding the rater-friendly rubric, because the rater-friendly rubric included specific, quantifiable bands. Raters stated that the conventional rubric was more difficult to use when scoring than the rater-friendly rubric, mainly because the rows in the conventional rubric contained too many descriptors. For example, in the conventional rubric, the "assignment prompt and content development" band include descriptors such as "arguments on all sides of the issue are identifiable, reasonable, and sound" which can arguably be less easily quantifiable.

The rater-friendly rubric seemed to be more reliable than the conventional rubric according to qualitative and quantitative data. Quantitative results revealed that there was more consistency in rating in the rater-friendly rubric (Inter-rater agreement of .95 as opposed to the conventional rubric's .86). While both rubrics would be reliable according to Stemler's (2004) acceptability value of .8., only the rater-friendly rubric's ICC coefficient of .72 meets the definition of sufficiently reliable alpha (.70) according to Brown, Glasswell, and Harland (2004).

Mean scores between the conventional rubric and the rater-friendly rubric were correlated, and produced a correlation coefficient of r=.41. See Figure 4. The scores given using both rubrics do not strongly correlate, which may mean that the rubrics measure different

constructs. This reinforces Atkinson and Ramanthan's (1995) claim that second-language writing follows a different approach and produces different results than first language writing. Using a rubric meant to assess first language writing may not be appropriate in a second language writing context. Rubrics that more closely represent the second-language writing process should be used.

The results of this study indicate that the rater-friendly rubric may be more appropriate than the conventional rubric to assess second-language writing in a first year composition course. This was likely because the rater-friendly rubric more closely matched classroom content and the writing prompt than the conventional rubric, indicating that second language writing instructors should create rubrics that assess what is taught and that closely match essay prompts to assign more valid and reliable scores.

Relevance to PIE and Second Language Learning

This study can help inform PIE English 105 instructors about how to create rubrics that are appropriate for their students' needs. Particularly, this study indicates that second language writing instructors should adapt materials used in mainstream courses to make them comprehensible and appropriate for ESL students. Additionally, developing assessment tools to fit the needs and expectations of ESL students can provide a more reliable picture of their progress in the course. This can inform future PIE English 105 classes about what to focus on in instruction.

References

Atkinson, D., & Ramanathan, V. (1995). Cultures of writing: An ethnographic comparison of L1 and L2 university writing/language programs. *TESOL Quarterly*, 29(3), 539-568.

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

Becker, A. (2010). Examining rubrics used to measure writing performance in US intensive English programs. *The CATESOL Journal*, 22(1), 113-130.

Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, 9, 105-121.

East, M., & Young, D. (2007). Scoring L2 writing samples: Exploring the relative effectiveness of two different diagnostic methods. *New Zealand Studies in Applied Linguistics* (13), 1-21.

Erdosy, M.U. (2004). Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions (40). Retrieved from Educational Testing Service, https://www.ets.org/Media/Research/pdf/RR-03-17.pdf.

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly*, *28*(2), 414-420.

Huot, B., & O'Neill, P. (2009). *Assessing Writing: A Critical Sourcebook*. Macmillan.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*(1), 36-62.

Miller, M. D., Linn, R., & Gronlund, N. (2013). *Measurement and assessment in teaching.* (11th ed.). New York, NY: Pearson.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4).

Yol, Ozge (*2015, April). Grading behavior of ENG 105 teachers and their use of rubrics.* Paper presented at the L2 Writing Club, Flagstaff, AZ.

Appendix
"Rater-Friendly" Rubric

Writing Project #3: Informational Argument Paper Rubric (15% of grade = 150 points)

Levels of achievement	Excellent	Good	Needs Work
Introduction 15	Introduces the topic and connects to thesis statement 15	Does not clearly introduce topic or connect to thesis statement 12	Introduction is off topic 9
Thesis Statement 5	Mentions the essay topic (technology in the classroom) and introduces subtopics (sides) that will be used in the essay (in the correct order) 5	Mentions the topic and two of the subtopics (sides) that will be used in the essay 3	Thesis does not mention the topic or 'sides' of the argument 1
Theme One 15	Both FOR and AGAINST are described with relevant and accurate citations (paraphrases or quotes) 15	Only FOR or AGAINST are described using sources 10	FOR and AGAINST are not clearly described 5
Theme Two 15	Both FOR and AGAINST are described with relevant and accurate citations (paraphrases or quotes) 15	Only FOR or AGAINST are described using sources 10	FOR and AGAINST are not clearly described 5
Theme Three 15	Both FOR and AGAINST are described with relevant and accurate citations (paraphrases or quotes) 15	Only FOR or AGAINST are described using sources 10	FOR and AGAINST are not clearly described 5
Conclusion 10	Recasts the thesis statement and restates the main ideas of the essay. Ends the essay with a larger idea. 10	Meets two of the following requirements:(a)recasts the thesis statement, (b) restates the main ideas of the essay, (c)ends the essay with a larger idea. 8	Conclusion does not recast the thesis statement, restate the main ideas, and end with a larger idea. 6
Neutrality 15	All sides of the issue are discussed without the author directly taking a side; hedging is used to avoid bias 15	Meets one of the following requirements: (a) all sides of the issue are discussed without the author directly taking a side, (b) hedging is used to avoid bias 12	Author clearly takes a side and does not use hedging to avoid bias 9
Visual Aids 5	One table, chart, or illustration supports a main point and is explained in the text; visual is formatted and cited according to APA 5	Meets one of the following requirements: (a) One table, chart, or illustration supports a main point and is explained in the text, (b)visual is formatted and cited according to APA 3	One table, chart, or illustration (e.g., visual) does not support a main point, is not explained in the text, and is not formatted and cited according to APA 1
Connections between ideas 15	The ideas throughout the essay are connected to each other using transition words 15	Half of the ideas in the essay are connected to each other using transition words 12	The ideas throughout the essay are not connected to each other 9
Formatting 10	APA formatting is followed 10	0-5 errors per page 8	10 or more errors per page 6
Grammar & Punctuation 5	No errors 5	0-5 errors per page 3	10 or more errors per page 1
Academic Word Choice 10	Uses academic language 10	Uses some academic language and some informal language 7	Word choice interferes with meaning 5