Influences of Rater Variables on College Japanese L2 Writing Assessment

SHIBATA Setsue
California State University, Fullerton
sshibata@fullerton.edu

日本語教育における大学生の作文はどう評価されるか：評価者の特徴とその影響に関する研究

柴田 節枝
カリフォルニア州立大学フラトン校

Abstract:

Although the Japanese L2 writing assessment is critical in evaluating a student's Japanese language skills, there has been little research into the effect of raters' background on Japanese L2 writing assessment. The present study examines the effects of four variables of the rater's background; academic specialty, teaching experience, attitudes toward composition in general, and attitudes toward the composition he/she is scoring as part of the College students' Japanese L2 writing assessment. Twenty-one college Japanese instructors assessed 15 compositions on both holistic and analytic scales. On the basis of their academic specialty, they were divided into 3 groups: Linguistics, Literature/Asian Studies and Education. The results show that academic specialty, teaching experience and attitudes toward composition in general are not the major factors affecting the raters' leniency, but their personal preferences for a writing they are scoring are one of the major factors that decrease inter-rater reliability of writing assessment. The study re-confirms the importance of multiple ratings to maintain fairness and accuracy in the writing assessment.

Introduction

The writing assessment, which is a critical part of language instruction, meets at least three purposes: program placement, monitoring student's progress, and accountability. Whatever the purpose, accuracy and reliability of rating are the key factors in an accurate assessment, and proof of an accurate and reliable assessment is consistency of score among raters (inter-rater reliability). Since a rater's judgement has always played an important role in the assessment of writing, adequate training and better specification of scoring criteria are crucial to minimize the raters' bias (Lumley, 2002). However, the substantial variation in

rater harshness (or leniency) that exists cannot be easily eliminated due to the nature of human beings (Carson & Carson, 1984; Lumley and McNamara, 1995; Kondo-Brown, 2002).

Although many previous studies investigated the factors which influence the inter-rater differences of L2 writing assessment (e.g., Cumming, Kantor, and Powers, 2002; Lumley, 2002; Song and Caruso, 1996), most of these studies are in the field of ESL/EFL. The purpose of this study is to investigate how a rater's variables such as educational background, attitude towards composition in general, and preference towards a student's writing that he/she is scoring influence his/her analytic and holistic scoring in the field of Japanese as a foreign or second language (L2).

Previous Studies

The assessment of writing proficiency is an essential part of L2 instruction, but is far more complex, challenging, and time consuming than with native speakers of the target language. Three types of rating scales are usually used in scoring a writing: analytic, primary trait, and holistic. Furthermore, each has a different purpose and focus in instruction and will provide different types of information to teachers and students (Cohen, 1994). Analytic scoring is considered the most appropriate when diagnostic and specific feedback is required, while holistic scoring is used to assess a student's overall performance particularly in case where only a limited time is available for assessment. Holistic scoring is often used in case of screening, placement, and accountability, e.g., to see if students have attained a relative expected level of proficiency. Holistic scoring is considered less reliable than analytic scoring, since it produces a single score in which the total quality of writing is not the sum of its components, but is viewed as a whole and tends to be more influenced by individual rater's characteristics than analytic scoring (Hamp-Lyons, 1995). Whatever the purpose of assessment and type of scoring used, the assessment needs to be performed accurately, consistently among raters, and effectively within the limited time available. Rater bias can be minimized by rater training

or experience, but is not likely to be eliminated completely due to an individual's unique characteristics (Kondo-Brown, 2002; Lumley, 2002).

Previous studies have been conducted focusing on the raters' characteristics on L2 writing assessment, particularly in the field of ESL/EFL. Song and Caruso (1996) compared ESL faculty and English faculty regarding the results of holistic and analytic evaluations of college students' essays written by non-native and native English speakers. They found no significant difference between the two groups of faculty on analytic rating, but found that the English faculty was more lenient on holistic rating than the ESL faculty. The study also found that with more experience in the writing assessment, raters became more lenient in their holistic evaluation. Lumley (2002) investigated the process by which raters make their scoring decision and found that the rating was heavily influenced by the individual intuitive impression of the text obtained when a rater first read it. Cumming, Kantor, and Powers (2002) found that the two groups of ESL/EFL teachers and English teachers for native English speakers used similar decision-making behaviors while assessing the TOEFL essays. They also found that the ESL/EFL teachers focused on language rather than on rhetoric and ideas, while the English teachers were more likely to focus on rhetoric and ideas in their overall assessment.

There are a number of studies on rater variables in the field of writing assessment of Japanese as L2. Among them Kondo-Brown (2002) investigated how judgments of raters were biased, focusing on the interaction of raters and types of writing. She analyzed the data of rating scores of college Japanese L2 writings rated by three raters using FACETS program. She found that all raters had their own unique bias patterns regardless of their relatively similar language and professional backgrounds. Her finding supports the necessity of multiple ratings even with the reliable assessment procedure. Tanaka, Tsubone, and Hajikano (1998) examined the differences between Japanese L2 teachers and non-teachers of native Japanese speakers regarding how the groups evaluate L2 Japanese compositions using analytic scoring. The ANOVA results indicated that

Japanese L2 teachers weighted on the content and the language use, and non-teachers weighted on the content and the accuracy, especially on the particles. It was also reported that teachers scored more leniently overall than non-teachers did.

As for the writing assessment of Japanese as L1 (*Kokugo*), Ishida and Mori (1985) found that there was a significant difference between elementary school teachers and college students regarding how they assessed the compositions of Japanese elementary school children. According to their study, teachers focused on language use while college students paid more attention to clearness of the theme. The study concluded that teachers' assessment reflected their own educational point of view, a bias that did not apply to the other group. Kajii (2001) investigated how the raters' psychological factors affect their assessment of writing. He analyzed the data obtained on 21 elementary school teachers in Japan, and found that the rater's personal preference towards a writing that he/she is rating is closely associated with a higher score. According to the report of Kokuritsu Kokugo Kenkyuujo [National Japanese Language Institute] (1978), elementary school children whose homeroom teachers' specialty is Japanese are more likely to have favorable attitudes towards composition. The study implies that the teachers' preferences and attitudes towards writing may influence their children's preferences and attitudes toward writing. Brown (1995) examined the effect of the raters' background on the oral assessment of Japanese using the Japanese Language Test for Tour Guides. She compared the results of oral assessment taken by 51 test candidates rated by the three groups of raters, based on their occupational background, i.e., a group that has guiding experience only, a group that has teaching experience only, and a group that have both guiding and teaching experiences. She also compared the differences between two groups of raters, a group of native speakers and a group of near-native speakers of Japanese. The results showed no significant difference among three groups of the occupational background regarding the consistency of rating, and showed no significant difference between native and near-native

raters' groups regarding harshness/leniency of rating.

As mentioned earlier, there is only a limited number of studies focusing on rater-variables on assessment of Japanese L2 writings. This study further examines the rater variables that may influence assessment of Japanese L2 writings.

The Study

Research Questions

In this study, the following questions are addressed.

1. Are raters of a particular academic background more harsh/lenient in assessing Japanese L2 writings?
2. Are raters with more experience teaching more harsh/lenient in assessing Japanese L2 writings?
3. Are raters' personal preferences/attitudes toward composition in general associated with harshness/leniency of their ratings?
4. Are raters' preferences toward a composition that they are scoring associated with harshness/leniency of their ratings?

Participants (raters of Japanese L2 writings)

Participants were 21 native Japanese language instructors who teach Japanese as L2 at the post-secondary level. All participants had at least a Master's degree, and their academic backgrounds were the following: Seven of 21 were Linguistics, eight were Education (Foreign Language Education, TESOL, and Second Language Acquisition), six were Literature, and four were Asian Studies. The range of their ages is from 26 to 58, and the average years of teaching Japanese as L2 is 10.1 years. They were asked to answer a variety of survey questions including length of experience teaching, educational background, personal preference for writing (1-5 scale where 5 is the highest). They were also asked to assess 15 compositions by both holistic and analytic methods.

Compositions

The compositions were written by 15 college students who studied Japanese as L2 and were in the second year or higher Japanese language classes. The students who agreed to participate in the study were asked to write a composition in the classroom. They were given 30 minutes to write a composition. Compositions were descriptive in nature, and the students chose one topic from the following: 1. The most favorite/unfavorable experience in your life; 2. Introducing your hometown; and 3. Introducing your favorite book or movie.

Data

Independent variables:

1. Raters' academic background: This data was obtained from the survey questionnaire. Participants were divided into three groups based on their academic backgrounds of either Linguistics, Education or Literature/Asian Studies. Four teachers of Asian Studies background were in the same group with teachers of Literature background, since they claimed that they took more Literature courses than other areas.
2. Teaching experience: Number of years of teaching experience was obtained from the survey questionnaire.
3. The raters' attitudes toward composition in general: The data was obtained from the survey questionnaire. Participants registered their level of their attitudes toward composition on a 5-point scale where 5 was 'the most favorite'.
4. The raters' preference/attitudes toward a composition that they are scoring: Participants were asked their level of preference of each of the Japanese L2 compositions that they were scoring. Rating was based on a 5-point scale where 5 was 'the most favorite.'

Dependent variables:

Each participant assessed 15 Japanese L2 writings. He/she assessed each

composition by two types of scoring, i.e., holistic and analytic rating.

1. Holistic rating score: ACTFL writing scale (Breiner-Sanders, Swender, and Terry, 2002) was used for the holistic scale. Level of proficiency is converted to the numerical numbers, from 1 (Novice-low) to 10 (Superior).
2. Analytic rating score: A modified scale developed by Sasaki and Hirose (1999) was used for the analytic scale. Since the scale is primarily for Japanese L1 writings, rubrics that were considered as either not appropriate or not clearly stated for evaluating an L2 writing were eliminated. There are five assessment areas, i.e., Content, Organization, Language use and vocabulary, Mechanics/Accuracy, and Appeal to the readers. Each area has two to five specific rubrics, and the scale of each rubric is a 5-point scale, where 5 is 'the strongly agree.' The total score is the sum of the scores of all rubrics. The highest score is 95 (5 x 19 rubrics).

Results

Table 1

Table of Means and Standard Deviations of Ratings by the Raters' Specialty

Group	Linguistics		Education		Literature/Humanities	
	Mean	SD	Mean	SD	Mean	SD
Analytic (n=23)	60.85	6.18	59.00	4.20	60.87	4.98
Holistic (n=23)	6.86	1.07	6.67	1.03	6.70	0.93

The means and standard deviations of holistic and analytic rating scores that were conducted by raters of three different groups of specialties are shown in Table 1. As illustrated in the table, the group of raters with an Education background gave the students lower score (i.e., rated them more harshly) than the other two groups in both types of ratings. One way ANOVA was conducted for holistic rating and analytic rating separately to examine the differences among the groups' rating characteristics.

Table 2

Analysis of Variance on Raters' Specialty for Analytic and Holistic Ratings

	Sum of Squares	df	Mean Square	F	Sig.
Analytic	33.75	2	16.88	0.66	0.5
Holistic	0.28	2	0.14	0.15	0.87

Table 3

Summary of Means and Analysis of Variance for Analytic Scoring Rubric

Rubric	Ling. (n=7)	Edu. (n=8)	Lit./Human (n=10)	F
1. Content	8.86	8.50	9.30	0.31
1-1. Is the theme clear?	3.00	2.83	3.01	0.07
1-2. Is the theme supported by Sufficient factual information?	3.14	2.84	3.10	0.35
1-3. Is the content consistent with the title?	3.57	3.33	3.80	1.11
2. Organization	7.14	6.50	7.50	0.92
2-1. Are paragraphs appropriately formed?	3.57	3.33	3.80	0.66
2-2. Are all paragraphs logically connected?	3.56	3.17	3.70	0.86
3. Language Use and Vocabulary	12.00	11.83	12.20	0.06
3-1. Is word usage correct?	4.00	3.83	4.11	0.27
3-2. Are sentences sufficiently short?	4.14	3.82	4.30	0.69
3-3. Are sentences adequately connected with appropriate use of conjunction and demonstrative words?	4.42	4.00	4.40	0.93

4. Mechanics/Accuracy	17.86	17.17	16.60	2.32
4-1. Are the particles used correctly?	4.42	4.00	4.20	1.81
4-2. Is the verb/adjective conjugation used correctly?	3.85	3.83	3.60	0.57
4-3. Is the tense appropriately used?	4.86	4.67	4.40	1.35
4-4. Is the grammar correctly used (other than 4-1 – 4-3)?	4.57	4.50	4.50	0.04
4-5. Is kana/kanji character written correctly?	4.71	4.98	4.80	0.91
5. Appeal to the Readers	9.97	9.66	9.90	1.68
5-1. Is the handwriting neat?	4.98	4.83	4.96	0.25
5-2. Is there sufficient amount of kanji characters at this level?	4.86	4.98	4.90	0.39

As shown in Table 2, there was no significant difference among the three groups of raters in both the holistic rating scores and the analytic rating scores with regards to harshness/leniency. ANOVA was also conducted for each rubric of five areas of the analytic scoring to examine the differences among the groups. The result shows that there were no significant differences among groups for their harshness/ leniency in any rubric of analytic scoring (see Table 3).

Table 4

Correlation Coefficients Matrix Between Variables

Variables	1	2	3	4	5
1. Holistic	--	0.53**	-0.10	-0.04	0.63**
2. Analytic		--	0.18	-0.11	0.55**
3. Experience			--	0.53**	-0.01
4. Attitude				--	-0.07
5. Preference					--

Note. **p<.01.

Table 4 shows Pearson's Correlation Coefficients between variables. As shown in the table, four significant relationships were found; between holistic and analytic ratings, between holistic rating and preference of a writing that he/she scored, between analytic rating and preference of a writing that he/she scored, and between teaching experience and attitude towards writing in general.

Table 5

Summary of Simultaneous Regression Analysis for Analytic Rating

Variable	B	SE B	ß
Experience	-0.13	0.14	-0.20
Attitude	0.15	1.02	0.03
Preference	2.47	0.85	0.05**

Note. **p<0.01.

Table 6

Summary of Simultaneous Regression Analysis for Holistic Rating

Variable	B	SE B	ß
Experience	-1.64E-02	0.03	-0.14
Attitude	5.97E-02	0.18	0.07
Preference	0.53	0.15	0.63**

Note. **p<0.01.

Simultaneous regression analysis was conducted to see if the three independent variables, i.e., raters' teaching experience, attitude towards writing in general, and raters' preference of a writing that they scored influenced the holistic rating and analytic rating. As shown in Table 5 and Table 6, the results indicate that a rater's personal preference towards the writing that he/she scored was positively associated with both types of scoring while the rater's teaching experience and his/her attitude toward writing in general do not affect their rating scores on both types of rating. It was also found that there is a positive

correlation between the holistic ratings rated by the ACTFL scale and the analytic ratings using the modified Sasaki's analytic scale.

Discussion

The results of the current study revealed that the raters' area of academic specialty and number of years of teaching experience did not directly influence their harshness/leniency when rating the college Japanese L2 writings. Brown's study (1995) also did not find significant differences of oral assessment among the three groups of occupational backgrounds.

It is reported that differences in assessing writing were found between ESL/EFL teachers and English teachers (Song and Caruso, 1996), between Japanese L1 teachers and Japanese with no teaching experience (Tanaka, Tsubone, and Hajikano, 1998), between Japanese elementary school teachers and Japanese college students (Ishida and Mori, 1985). In these studies, raters' teaching background, i.e., either L1 or L2, was the focus, rather than their specialty. The participants in the current study are from three different academic specialties but all have L2 teaching experience. Results show that it was not the raters' academic specialty but the raters' teaching background, i.e., either trained as L2 teacher or not, that influenced their rating characteristics. According to Song and Caruso (1996), ESL teachers with longer experience assessing writing skills seem to be more lenient in their holistic evaluation. In this study, teaching experience did not have a significant positive relationship with their harshness/leniency in their rating. It is possible that longer teaching experience does not automatically mean longer assessment experience. However, since few studies have been conducted on teachers' specialty and writing assessment experience as a rater-variable on Japanese L2 writing assessment, further studies are needed to verify the results of this study.

It is also found that the raters' attitudes toward writing did not influence their ratings on both holistic and analytic scales. As Kokuritsu Kokugo Kenkyuujo suggested (1978), the instructors' positive attitudes toward writings are usually a factor that promotes the students' motivation and desire to learn

Japanese, but the current study indicates it did not affect their assessment directly. Whether attitudes toward writing are registered as 'the most favorite' or not for the raters was not a significant factor in their harshness/leniency of the assessment of Japanese L2 writings.

It was also revealed in this study that raters' preferences toward a writing they scored affected their ratings, i.e., raters tended to give a higher score to the writing that they felt the most positive about. In other words, personal preference toward a writing they scored makes the assessment rather subjective, and as a result can be a factor of a rater-bias. The study supports Lumley (2002), who concluded that rating is heavily influenced by the complex intuitive impression of the text obtained when the raters first read it. What makes the raters have a positive feeling/attitudes toward a writing is different from individual to individual; some weigh on rhetoric of the text, some on the theme and content, and others on handwriting and neatness. Such possible interactions between a composition and a rater is beyond the scope of the current study, but is required to be explored in future research.

Although a causal relationship is unknown, it was found that teachers with longer teaching experience tended to have positive attitudes toward writings in this study. Teachers who like to "write" tend to stay longer in the teaching position of a language, or they develop a more positive attitude toward writing as their teaching careers become longer. This should also be verified in future research with a much larger sample.

The purpose of this study was not to examine the reliability of the assessment tools, but intra-rater reliability was found between two types of rating scales, the ACTFL rating scale and the modified Sakaki's analytic rating scale.

Conclusion and Implications

Assessment is a critical part of Japanese language instruction, and it requires accuracy and consistency. Adequate rater-training and clear criteria of rating scale can eliminate biases among raters, but it is still not possible to obtain perfect accuracy and consistency because the assessment is conducted by a

human being. All raters have their own unique bias patterns, and such patterns and factors that make individual differences in assessment are very complicated and variable.

In this study, it was found that raters' academic background, number of years of teaching experience, and their attitudes toward writing were not the factors influencing their rating characteristics, i.e., harshness/leniency on Japanese L2 writing. It was also found that the raters' attitudes toward a writing that they are scoring makes the assessment rather subjective, i.e., raters tend to score higher when they have a positive feeling toward the writing they are scoring.

It can be concluded that the factors that cause rater-bias are more individualized and complex in nature, associated with the individual and their personality rather than the writing that they are scoring. Such mechanism of interactions between raters and writing texts should be explored in future research. Further exploration of similarities and differences among raters on assessment will likely yield a better understanding of the nature of assessment, and a more complete accurate understanding of the phenomenon of assessment in Japanese L2 writing process.

Acknowledgments

The author wishes to acknowledge the Faculty Development Center of California State University, Fullerton, which provided funding for this study.

References

Breiner-Sanders, K. E., Swender, E., & Terry, R. M. (2002). Preliminary proficiency guidelines – Writing revised 2001. *Foreign Language Annals, 35,* 9-15.

Brown, A. (1995). The effect of rater variables in the development of anoccupation-specific language performance test. *Language Testing, 12,* 1-15.

Cason, G. J. & Cason, C. L. (1984). A deterministic theory of clinical performance rating. *Evaluation and the Health Professions, 7,* 221-247.

Cohen, A. D. (1994). *Assessing language ability in the classroom.* Boston, MA: Heinle & Heinle.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86,* 67-96.

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly, 29,* 759-762.

Ishida, J. & Mori, T. (1985). Shougakusei no bunshou-hyougen-ryoku nohattatsu-teki henka [Changes in the development of writing skills among elementary school children]. *Hiroshima University, Department of Education Bulletin, 33,* 125-131.

Kajii, Y. (2001). Jidou no sakubun wa dono you ni hyouka sareru noka? [How do teachers evaluate elementary school children's composition?]. *Shinrigaku Kenkyuu* [Journal of Educational Psychology], *49*, 480-490.

Kokuritsu Kokugo Kenkyuujo [National Japanese Language Institute]. (1978). *Jidou no hyougen-ryoku to sakubun* [Children's self-expression and composition skills]. Tokyo: Tokyo Shoseki Publishing.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19,* 3-29.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19,* 246-276.

Lumley, T. & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12,* 54-71.

Sasaki, M. & Hirose, K. (1999). Development of an analytic rating scale for Japanese L1 writing. *Language Testing, 16,* 457-478.

Song, B. & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing, 5,* 163-182.

Tanaka, M., Tsubone, Y., & Hajikano, A. (1998). Daini-gengo toshite nonihongo ni okeru sakubun hyouka kijun: Nihongo kyoushi to ippan nihon-jin no hikaku [Evaluation criteria for Japanese L2 writing: A comparison between Japanese language teachers and the general public]. *Nihongo Kyouiku* [Journal of Japanese Language Education], *96,* 1-12.