Development of an analytic rating scale for Japanese L1 writing

Miyuki Sasaki *Nagoya Gakuin University* and **Keiko Hirose** *Aichi Prefectural University, Japan*

The present study developed an analytic rating scale for Japanese university students' first language (L1) expository writing. Research proceeded in four stages. In Stage 1, we devised a questionnaire to investigate Japanese L1 teachers' criteria for evaluating expository writing. A total of 102 teachers responded to the questionnaire, and rated the 35 related descriptions according to their importance. In Stage 2, based on the questionnaire results, we either eliminated or reorganized these 35 descriptions under six analytic criteria:

1) Clarity of the theme;
2) Appeal to the readers;
3) Expression;
4) Organization;
5) Knowledge of language forms; and
6) Social awareness.

Then, in Stage 3, we attempted to investigate possible weighting of these six criteria. We asked 106 teachers to rate two to four compositions with varied profiles using both a holistic scale and the obtained analytic scale. The results showed that the explanatory power of each criterion can vary from composition to composition, and thus we concluded that the six criteria should have equal weighting. Finally, in Stage 4, we pilot-tested the obtained scale using a new set of 69 compositions. The results indicate that the present scale is both valid and reliable, and that it is superior to a traditional analytic scale in capturing such composition qualities as appeal to the readers and social awareness.

I Introduction

The present study was motivated by a larger study (Sasaki and Hirose, 1996), which investigated the relationship between Japanese university students' first (L1) and second (L2) language writing ability. In a pilot study (Hirose and Sasaki, 1994), we discovered that there was no standard scale for rating Japanese L1 compositions that could be used in research and teaching contexts ('Donna Kantende Sakubun o

Address for correspondence: Miyuki Sasaki, Associate Professor, Faculty of Foreign Studies, Nagoya Gakuin University, 1350 Kamishinano-cho, Seto-shi, Aichi-ken, 480–1298, Japan; e-mail: f43689g@nucc.cc.nagoya-u.ac.jp

Language Testing 1999 16 (4) 457–478 0265-5322(99)LT180OA © 1999 Arnold

Hyoukasuruka [Evaluating composition in Japanese as L1]', 1992). Due to the lack of standards, researchers either have created an ad-hoc rating system (e.g., Carson *et al.*, 1990), or have used translated/adapted versions of the existing English L2 rating scales (e.g., Pennington and So, 1993). Although such 'make-shift' treatments may have been appropriate for the cases in question, they may have overlooked certain aspects of the true L1 readers' judgement, thus losing part of their validity (cf. Carpenter *et al.*, 1995). Considering the recent 'appeal to "real-life performance"' (Bachman, 1990: 304) in language testing, the existence of valid and reliable rating scales for directly assessing both L1 and L2 writing quality seems especially necessary when we compare students' L1 and L2 writing ability.

Thus, we designed the present study to develop a scale for rating Japanese university students' L1 expository writing. An analytic or 'profile' type of scale, which assesses each composition according to multiple dimensions as opposed to a holistic scale giving only one score (Huot, 1990), was chosen because:

1) it is comparable with the ones used for students' English L2 writing ability (the Profile Scale of Jacobs *et al.*, 1981) in the larger study;
2) it has proved to be more reliable than other types of scales (e.g., holistic scoring, primary trait scoring (see Jacobs *et al.*, 1981; Hamp-Lyons, 1991, 1995; Huot, 1990); and
3) it provides useful diagnostic information not found in other methods (Jacobs *et al.*, 1981; Raimes, 1990; Hamp-Lyons, 1991).

1 *Development of analytic scales for writing in other languages*

Analytic scoring is one of the main procedures for directly assessing writing ability (Huot, 1990). It was originated by Diederich *et al.* (1961), who factor analysed 53 English L1 readers' remarks about 300 college compositions. They identified five major traits that the readers tended to value when rating the compositions (ideas, forms, flavour, mechanics and wording), and devised the first analytic scale for writing in English as L1 (Diederich, 1974). Although Diederich *et al.* found that each reader put different emphases on these five traits, in the actual scale they assigned 10 points to the first two traits, and five points to the others because the former traits were stressed in the writing courses where the scale was used. Diederich explained that, 'This weighting had no basis in research, but it seemed reasonable to give extra credit for the qualities these teachers wished to emphasize' (1974: 54). The scale has since been criticized for mechanical or unquantifiable content, and low reliability (partly due to lack of proper rater training; see Hamp-Lyons, 1991), but it still has significance in pioneering the methodology of developing analytic rating scales for compositions.

Another noteworthy effort in the field of L1 writing was made by the International Association for the Evaluation of Educational Achievement (IEA) (Gorman *et al.*, 1988). They developed a common scoring scheme and corresponding scales for comparing high-school students' compositions written for 14 different tasks across 14 different countries. Their final rating scheme was obtained through long and careful research efforts including reviewing the relevant literature, synthesizing the international reading teams' comments, and conducting several pilot sessions. The IEA's general scoring scheme consisted of seven core components. These were:

1) quality and scope of content;
2) organization and presentation of content;
3) style and tone;
4) lexical and grammatical features;
5) spelling and orthographic conventions;
6) handwriting and neatness; and
7) response of the rater.

Each of the components was elaborated with detailed guiding descriptions according to the different demands of each writing task (e.g., narrative). The IEA study was important because it represented a true effort to establish a modern sense of construct validation for the use of their scales, by first explicitly defining the construct to be evaluated (i.e., students' writing ability in the given context), and then collecting evidence to 'support the adequacy and appropriateness of inferences and actions based on test score' (Messick, 1993: 13).

In the field of L2 writing, Jacobs *et al.*'s (1981) writing profile was one of the first attempts to develop an analytic type of scale. The profile was originally developed to assess a large number of compositions that were part of an English L2 proficiency test battery (the Michigan Test battery). In search of an evaluation system that could achieve high reliability with 'a large number of relatively inexperienced readers' (p. 33), Jacobs *et al.* developed a five-component scale (content, organization, vocabulary, language use and mechanics). These five components were derived from the already established scheme for rating the compositions written by international students at Texas A. & M. University, but were later elaborated and refined into the present form based on the reviews of previous literature and several pilot-study results. It is noteworthy that Jacobs, *et al.* determined the weighting of each component according to what they believed composition teachers should value most in students' composition (i.e., communicative effectiveness; thus, content and organization) rather than what the teacher actually emphasized most when grading the students' compositions (e.g., mechanics). Although such an approach (starting with a certain construct – i.e., communicative effectiveness – to be measured) is an important step for constructing a valid test,[1] it is regrettable that their decision on the weighting of each criteria was not completely supported by statistical evidence.

More recently, Hamp-Lyons (1991) advocated 'multiple-trait scoring', which is similar to analytic scoring in that it provides more than one trait score, but places more emphasis on the context in which the writing takes place. Thus, 'multiple trait scoring procedures are developed on-site for a specific purpose with a specific group of writers, and with the involvement of the readers who will make judgments in the context' (p. 248). Hamp-Lyons demonstrated this process by describing in detail her own experience of developing multiple-trait instruments. For example, when she developed the scale for assessing the incoming undergraduate students' composition at the University of Michigan, she started by collecting the target readers' (faculty teaching composition and other subjects, and senior teaching assistants) definitions of 'good student writing' in the given context. Then in the subsequent 'recursive' (p. 259) development processes, Hamp-Lyons had to 'reconcile and incorporate the implicit constructs or definitions of good writing that emerged from the data from the different groups' (p. 259). As a result, she created the 'Michigan Writing Assessment Scoring Guide' with three components with detailed descriptions for each of six levels; the three main components are:

1) ideas and arguments;
2) rhetorical features; and
3) language control.

Hamp-Lyons reported that the new scoring method was well responded to by everyone involved, including the composition readers, faculty advisors and students themselves, because all of them could now share the same scoring guide to anchor their discussion.

Considering the merits and demerits of the procedures employed

[1] In fact, Jacobs *et al.* (1981: 73) clearly defined the construct they intended to measure as 'an ESL writer's ability to process discourse for effective communication', and showed evidence for the validity of their scale.

for developing those previous analytic type scales for writing in other languages, we decided to proceed according to the following four stages to reach the final form of the scale.

Stage 1: Development of a questionnaire to investigate Japanese L1 (writing) teachers' criteria for evaluating expository writing;
Stage 2: Analysis of the questionnaire results and creation of a first draft of the scale;
Stage 3: Investigation of possible weighting of the obtained evaluation criteria and preliminary validation check of the scale;
Stage 4: Pilot testing the obtained scale.

Before we launched into these stages, however, we defined the test task for which the target scale was going to be used as 'Japanese university students' L1 expository compositions'. An expository composition or 'SETSUMEIBUN' in the present study was defined as 'a text written for the purpose of explanation and/or persuasion (translation ours)' following Ueyama and Morita's (1990) definition.[2]

II Studies

1 Development of a questionnaire to investigate Japanese L1 teachers' criteria for evaluating expository writing

In Stage 1, following Converse and Presser's (1991) methodology for writing survey questions, we developed a questionnaire to investigate Japanese L1 teachers' criteria for evaluating expository writing. We judged that a questionnaire would make it easier for us to gather much more information from the participants (i.e., busy high-school teachers) than a free writing form.

First, we asked two Japanese L1 writing experts (university professors) to survey the previous literature on what has been valued when Japanese L1 teachers evaluate expository writing (e.g., Komatsu, 1976; Shiojiri, 1978; Kinoshita, 1981; Kokugo Kyoiku Kenkyusho [National Language Education Research Institute], 1988; Ueyama and Morita, 1990; Oouchi, 1995). Based on the result of the survey, the two experts proposed five major areas (i.e., expression, organization, content, appeal to the readers and social awareness) with 35 corresponding descriptions that Japanese L1 teachers may possibly use as internal evaluation criteria. The 35 descriptions were intended to reflect the two main schools of thought that have been influential

[2] The description of such a test context followed the procedure suggested in the draft version of Bachman and Palmer (1996), which was used as a textbook for Bachman's class at the TESOL 1994 Summer Institute, Iowa State University.

in the Japanese L1 composition classrooms over the past 50 years (Oouchi, 1995): the Kokugoka Sakubun (Expression) school and the Seikatsu Tsuzurikata (Life-Experience Writing) school.[3]

In cooperation with the two Japanese L1 writing experts mentioned above, we then revised and refined the 35 descriptions into questionnaire entry forms. The questionnaire asked the participants to rate the descriptions on a five-point Likert scale according to the importance they would attach when evaluating expository compositions (Appendix 1). After the first draft was made, we further revised the wording of each entry and the overall organization based on the results of several pilot trials and follow-up interviews with the participants (university and high-school Japanese L1 teachers).

2 Analysis of the questionnaire results and creation of the first draft of the scale

The final version of the questionnaire was sent out to 200 high-school teachers in different parts of Japan. High-school teachers were selected as participants because Japanese writing (Kokugo Hyougen) is taught as an important part of the high-school curriculum (Ministry of Education, Science and Culture, Government of Japan, 1989), and university students' compositions, the target of our scale, can be regarded as the end product of such nationally uniform instruction. Moreover, it was difficult to find university teachers who regularly read and assess students' L1 compositions because Japanese L1 writing is rarely taught at Japanese universities (Kinoshita, 1990).

A total of 102 teachers (57 men and 45 women; mean age: 33.3 years) with an average of 10.4 years of teaching experience from 12 prefectures responded. Because the means of the 35 description variables of the questionnaire indicate how much they were valued by the participants, we decided to eliminate those that were rated relatively low on the Likert scale from the final inclusion in the rating scale. To be cautious for the preliminary version of the scale, however, we only dropped the 11 descriptions (Items 1, 4, 6, 11, 13, 16, 18, 29, 31, 33, 35) with means below 3.51, i.e., the grand mean (3.67) minus 1.96 standard error of measurement (0.082).

[3] The Expression school emphasizes well-planned and systematic curricula guided by the Course of Study (Ministry of Education, Science and Culture, Government of Japan, 1989). This school includes those practitioners who have been influenced by the academic rhetoric instruction in the United States (e.g., Kinoshita, 1981). In contrast, the Life-Experience Writing school focuses on students' cognitive and personal growth through writing (Kitagawa and Kitagawa, 1987). It regards the process of writing as a means of self-discovery rather than as just a means of acquiring writing skills.

The remaining 24 descriptions were reorganized under the following six criteria:

1) Clarity of the theme;
2) Appeal to the readers;
3) Expression;
4) Organization;
5) Knowledge of language forms;
6) Social awareness.

Following several intensive discussions with the two L1 Japanese experts mentioned above, we renamed the category of 'Content' as 'Clarity of the theme', emphasizing the most highly rated among the six relevant descriptions (numbers 22 and 23). The experts also recommended that the remaining eight descriptions for the 'Expression' category should be divided into two, i.e., one concerned with more functional aspects of expression (e.g., cohesion, coherence) and one concerned with more formal aspects (e.g., orthographical notation, punctuation, language use) if they were to be used as evaluation criteria. The experts argued that giving the users separate scores for these two different aspects would be more useful for diagnostic purposes because these two aspects are often emphasized differentially in Japanese composition classes. Thus, we divided the original category of 'Expression' into 'Expression' (concerning more functional aspects) and 'Knowledge of language forms' (concerning more formal aspects). The names of the other three original criteria remained the same. Table 1 presents the descriptive statistics for the items selected for each of these six criteria. Mean, skewness and kurtosis values for these variables indicated relatively normal distributions (for the criteria of normality; see Sasaki, 1996).

Finally, these 24 reorganized descriptions were further revised so

Table 1 Descriptive statistics of the 24 description (item) variables selected for the six criteria ($N = 102$)

Criterion	Mean*	SD	Kurtosis	Skewness
1. Clarity of the theme [Items 19, 20, 21, 22, 23, 24**]	4.02	0.84	–0.04	–0.58
2. Appeal to the readers [Items 25, 26, 27, 28]	3.76	0.84	–0.61	–0.18
3. Expression [Items 5, 9, 10, 12]	4.11	0.75	–0.47	–0.43
4. Organization [Items 14, 15, 17]	3.90	0.83	–0.24	–0.38
5. Knowledge of language forms [Items 2, 3, 7, 8]	3.95	0.86	–1.00	–0.21
6. Social awareness [Items 30, 32, 34]	3.78	0.89	–0.59	–0.17

Notes: All values presented are the average of the selected variables for each criterion; * Each score can range from 1 to 5; **See Appendix 1 for the content of each item.

that they would be appropriate as descriptors for the six selected analytic criteria. In the obtained preliminary scale, each of these six criteria had equal weighting (10 points; see Appendix 2a and Appendix 2b for a translation).

3 Investigation of possible weighting of the six analytic criteria and preliminary validation check of the scale

In Stage 3 we attempted to determine the weighting of the six analytic criteria obtained in Stage 2 through several steps. Although we realized that some researchers have been critical of assigning different weights to such evaluation criteria (e.g., Hamp-Lyons, 1991), we still hypothesized that our six criteria might differentially contribute to the raters' overall judgment. As a by-product of this investigation, we could also check one aspect of the validity of the obtained scale, i.e., how much the selected six analytic scores together explained the raters' overall holistic judgement.

First, we collected 70 samples of expository (argumentative) compositions from first-year undergraduates (26 men and 44 women; mean age: 18.3) majoring in British and American studies.[4] An argumentative task (writing for or against the idea of married women working outside their home) was chosen because it had been the most popular among Japanese first-year university students who had discussed 10 different topics in English (Hirose and Kobayashi, 1991). The collected compositions were rated by two experienced Japanese L1 teachers using the six criteria obtained in Stage 2. When the raters' scores for one criterion were different by more than eight points, we employed a third rater (another experienced and trained Japanese L1 teacher).[5] Then, among the 70 compositions, we selected the two with the greatest variations among the six criterion scores (Compositions 1 and 2; see Table 2) so that we could effectively investigate the relationship between these scores and the raters' overall judgement. We sent these two compositions to 200 Japanese L1 (mainly high-school) teachers in different parts of Japan. Only two out of the 70 compositions were sent to these potential participants because, according to some previous interviews, two compositions were the maximum that most of those busy teachers could find time to rate.

For these two compositions, the teachers were asked to give analytic scores using the obtained scale, and holistic scores according to the following scale:

[4] These 70 compositions were also used for the larger study (Sasaki and Hirose, 1996).

[5] These three raters were different from the experts who helped us with the Stage 1 and 2 studies.

Table 2 Different profiles suggested by the six analytic scores of compositions 1 to 4 used in Stage 3

Criterion (total possible* in brackets)	Composition			
	1	2	3	4
1. Clarity of the theme (20)	13	12	3	19
2. Appeal to the readers (20)	7	14	3	17
3. Expression (20)	9	9	14	18
4. Organization (20)	6	10	12	19
5. Knowledge of language forms (20)	14	11	17	20
6. Social awareness (20)	7	15	3	18
7. Total score (120)	56	71	52	111

Note: * Sums of the two raters' scores.

- 10–9 Very good;
- 8–6 Good;
- 5–3 Fair;
- 2–1 Poor.

In order to avoid the influence of the analytic rating, the teachers were asked to give the holistic score before the analytic scores. The order of the two compositions was randomized. A total of 106 teachers (104 Japanese L1 high-school teachers and two elementary-school teachers specializing in L1 writing; 69 males and 37 females; mean age: 37.0; mean years of teaching: 12.75) from 20 prefectures provided complete data. Furthermore, 33 out of the 106 teachers agreed to rate two additional compositions of different profiles (Compositions 3 and 4; see Table 3).

Next, the relationship between the six analytic scores and the holistic score for the same compositions were examined by regression analysis. As a preliminary step, the holistic score was regressed on the six analytic scores as the independent variables. This regression procedure was repeated for the ratings of Compositions 1 and 2 (106 raters) and Compositions 3 and 4 (33 raters); $F(6, 99) = 36.40$, $p < .001$ for Composition 1; $F(6, 26) = 33.30$, $p < .001$ for Composition 2; $F(6, 26) = 8.08$, $p < .001$ for Composition 3; and $F(6, 26) = 13.83$, $p < .001$ for Composition 4. The six analytic scores together explained significant portions of the holistic score variances in all four compositions (the coefficient of determination (R^2), was 0.69 for Composition 1, 0.67 for Composition 2, 0.65 for Composition 3 and 0.76 for Composition 4). That is, over 65% of the raters' holistic scores for all four compositions was accounted for by the six analytic scores, which provides concurrent evidence to support the validity of the analytic scale's score interpretation.

Finally, we examined the individual contribution of the six analytic scores by calculating the product–moment correlation coefficients for the relationship between each of these six scores and the holistic score. The squares of these coefficients indicate the proportion of the holistic-score variance explained by each of these six scores. This additional analysis was necessary because the six independent variables (the six analytic scores) in the above regression analyses were highly correlated with each other for all four compositions. In such cases, the independent variables' contributions to the dependent variable (the holistic score) overlap, and it is impossible to examine the 'total' contribution of each independent variable through the results of one comprehensive multiple-regression analysis only (Tabachnick and Fidell, 1989).

The resulting correlation coefficients for the relationship between each of the six analytic scores and the holistic score, and the squares of these coefficients, were statistically significant ($p < .01$) for all four compositions, which supports the validity of including these six criteria in the scale. However, the squares of the correlation coefficients for the same analytic scores (i.e., their individual contributions to the holistic score variance) sometimes varied greatly across four compositions (Table 3). For example, the square of the correlation coefficient for the 'Organization' score of Composition 3 (0.21) was only about one third of that for the 'Organization' score of Composition 4 (0.62). These results indicate that the raters differentially weighted each criterion when they responded to different compositions. Based on these results, we decided against the idea of weighting the criteria, deciding, instead, to weight each criterion equally (10 points each) as in the original scale. If the explanatory power of each criterion can vary from composition to composition as in the present

Table 3 Squares of the correlation coefficients for the relationship between the analytic and holistic scores of the four compositions

Analytic criterion	Composition			
	1	2	3	4
1. Clarity of the theme	0.44	0.39	0.43	0.62
2. Appeal to the readers	0.57	0.45	0.54	0.48
3. Expression	0.21	0.45	0.18	0.53
4. Organization	0.38	0.55	0.21	0.62
5. Knowledge of language forms	0.25	0.22	0.21	0.34
6. Social awareness	0.38	0.31	0.14	0.49

study, it might be wise to stay with equal weighting as suggested by Hamp-Lyons (1991), or to determine the weighting of each criterion according to the given instructional emphasis as in Diederich *et al.* (1961).

For the final version of the rating sheet, we further revised the wording of the descriptors, considering the participating teachers' comments and suggestions returned with their ratings.

4 Pilot testing the obtained scale

In the last stage of the present study, we pilot-tested the final version of the obtained analytic scale. The reliability of the scale was estimated by calculating the product–moment correlation coefficients for the two raters' scores (Bachman, 1990), whereas the validity was examined by comparing the usefulness of the obtained scale with that of a traditional one (cf. Carpenter *et al.*, 1995 for the methodology for comparing the qualities of two tests).

a Procedure: First, we obtained 69 samples of expository compositions with the same topic used in Stage 3 (writing for or against the idea of married women working outside their home) from a new group of participants: 69 first-year undergraduates (19 men and 50 women; mean age: 18.2) majoring in British and American studies. Their characteristics fit those of the population targeted by the obtained scale. Then we asked two Japanese L1 composition teachers to rate the 69 compositions using the obtained scale (henceforth, the New Scale).[6] Six months later, we again asked the same two teachers to rate the same 69 compositions using the translated version of Jacobs *et al.*'s (1981) writing profile (henceforth, the Old Scale), which has traditionally been used to measure students' Japanese as L1 or L2 writing ability (e.g., Hirose and Sasaki, 1994; Pennington and So, 1993). We waited for six months to minimize the possible influence of the first rating.

For both rating sessions, the two raters were carefully trained and normed so that they could properly use each scale. When their scores differed by more than 7 points for one criterion, attempts were made to make the differences smaller according to Jacobs *et al.*'s (1981) recommended procedure. Furthermore, in order to avoid possible order effects, the raters in both sessions rated the compositions in opposite orders. The sum of the two raters' scores was used for the

[6] One of these teachers was different from those who participated in the Stage 3 study.

subsequent analyses (except for the investigation of the inter-rater reliability).

b Descriptive statistics: The descriptive statistics for the New Scale are presented in Table 4. Skewness and kurtosis values for the six analytic scores indicated relatively normal distributions. All of the six scores had relatively wide ranges, indicating that they captured different levels of ability among the participants.

c Reliability: Table 5 shows the correlations between the two raters' scores on analytic scores as well as the total scores for the New Scale and Old Scale. The obtained coefficients for the New Scale

Table 4 Descriptive statistics of the six criterion scores for the New Scale (*n* = 69)

Criterion	Mean (total possible)	SD	Kurtosis	Skewness	Range
1. Clarity of the theme	13.77 (20)	3.38	–0.36	–0.59	13
2. Appeal to the readers	11.26 (20)	3.51	–0.71	–0.13	14
3. Expression	12.30 (20)	2.74	–0.23	0.58	12
4. Organization	10.93 (20)	2.96	–0.22	0.52	12
5. Knowledge of language forms	14.12 (20)	3.16	0.92	–1.04	14
6. Social awareness	12.39 (20)	3.53	–0.02	–0.53	16

Table 5 Correlation coefficients between the seven scores given by the two raters

New Scale	
1. Clarity of the theme	0.69
2. Appeal to the readers	0.68
3. Expression	0.53
4. Organization	0.56
5. Knowledge of language forms	0.60
6. Social awareness	0.63
7. Total score	0.84
Old Scale	
1. Content	0.73
2. Organization	0.51
3. Vocabulary	0.52
4. Language use	0.52
5. Mechanics	0.57
6. Total Score	0.80

ranged from 0.53 to 0.84, while those for the Old Scale were also high enough (0.51 to 0.80) to permit further analyses.

d Comparison between the results of the New and Old Scale ratings: The results of the New Scale rating were then compared with those of the Old Scale rating for the same 69 compositions. The total scores produced by the two rating systems correlated relatively highly (0.76), suggesting that the two scales measured a similar trait. However, there was a group of compositions for which the total scores produced by the New Scale and Old Scale were very different. We selected 10 such compositions whose total scores differed most between the New Scale and Old Scale, and compared them with the other 59 compositions. We found that these 10 compositions' New Scale percentage mean-total scores (53.4/120 = 44.5%) were much lower than their Old Scale percentage mean-total scores (140.7/200 = 70.4%). This created the difference between these 10 compositions and the other 59 because the latter group's New Scale percentage mean-total scores (65.3%) were not as drastically low as their Old Scale percentage mean-total scores (73.7%).

The difference between the old and new total scores for the 10 compositions was mainly caused by the fact that they received much lower scores for Clarity of the theme, Appeal to the readers and Social awareness than the other 59 compositions (Table 6). This finding indicates that, unlike the old scale, the new scale properly flagged those compositions that were especially weak in these less directly linguistic areas. Because most specifications for these criteria (especially those for Appeal to the readers and Social awareness)

Table 6 Comparison between the 10 students whose New-Scale total score was much lower than their Old-Scale total score and the other 59 students

Score (total possible in brackets)	The 10 students' percentage mean	The 59 students' percentage mean	Difference
Old Scale total (200)	70.4	73.7	3.3
New Scale total (120)	44.5	65.3	20.8
New Scale			
1. Clarity of the theme (20)	50.0	72.0	22.0
2. Appeal to the readers (20)	33.0	60.5	27.5
3. Expression (20)	52.5	63.0	10.5
4. Organization (20)	40.5	57.0	16.5
5. Knowledge of language forms (20)	54.5	73.5	19.0
6. Social awareness (20)	36.5	66.5	30.0

reflect a particularity of the Japanese L1 composition teachers' judging criteria, this finding warns against using a translated version of a rating scale for a non-intended language, which might have different judgemental values in assessment (cf. Hinds, 1987).

III Conclusions

The present study developed an analytic rating scale for Japanese university students' L1 expository writing. In the process, we tried to incorporate as many Japanese L1 composition teachers' internal judging criteria as possible into the rating specification. In the pilot trial where we used argumentative expositions, the new scale was both reliable and valid. However, further studies related to the use of this new scale are necessary to develop a better instrument. Numerous validation studies conducted for L1 English rating scales provide methodological guidance in this area (e.g., Huot, 1990). First, different types of expositions (e.g., explanation) should be used to test the applicability of the scale. Second, the possibility of applicability should be further sought by using the scale for different types of writing in different contexts (cf. Hamp-Lyons and Henning, 1991). For example, the scale might be useful without major revisions for high-school students' summaries or narratives. Third, the impact of using the scale in the classrooms should be reported by the users. For example, a favourable wash-back effect would further promote the use and improvement of the present scale. Finally, a manual for training the raters should also be developed to optimize the use of the scale.

Acknowledgements

An earlier version of this paper was presented at the 11th World Congress of Applied Linguistics, 1996. We thank those who contributed to the completion of the present study. Toshihiro Ueyama and Hirotaka Namba provided us useful information on teaching L1 writing in Japan, and helpful comments throughout the whole process of the present study. Masaru Sumida, Hitoshi Yayama, Kazuo Matsutomo and Jun Nakanishi, rated the Japanese compositions. Phillip R. Morrow and William J. Herlofsky provided valuable comments and suggestions on earlier versions of the present paper. We are also indebted to Toru Kinoshita, who helped us collect the data, and Yasuko Nogami, who helped us with the statistical procedures. Needless to say, the positions taken and any errors that may remain are solely our responsibility. The research in the present paper was funded by a 1995 academic-year research grant from Nagoya Gakuin University.

IV References

Bachman, L.F. 1990: *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L.F. and **Palmer, A.S.** 1996: Language Testing in Practice. Oxford: Oxford University Press.

Carpenter, K., Fujii, N. and **Kataoka, H.** 1995: An oral interview procedure for assessing second language abilities in children. *Language Testing* 12, 157–81.

Carson, J.E., Carrell, P.L., Silberstein, S., Kroll, B. and **Kuehn, P.A.** 1990: Reading-writing relationships in first and second language. *TESOL Quarterly* 24, 245–66.

Converse, J.M. and **Presser, S.** 1991: *Survey questions: handcrafting the standardized questionnaire*. Newbury Park, CA: Sage.

Diederich, P.B. 1974: *Measuring growth in English*. Urbana, IL: National Council of Teachers English.

Diederich, P.B., French, J.W. and **Carlton, S.T.** 1961: *Factors in judgments of writing ability*, Research Bulletin 61–15. Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction No. ED 002 172.)

Donna Kantende Sakubun o Hyoukasuruka [Which criteria should be used for assessing Japanese L1 composition?] 1992: *Kyouiku Kagaku Kokugo Kyouiku* [Educational Science of Teaching Japanese].

Gorman, T.P., Purves, A.C. and **Degenhart, R.E.,** editors, 1988: *The IEA study of written composition I: the international writing tasks and scoring scales*. Oxford: Pergamon Press.

Hamp-Lyons, L. 1991: Scoring procedures for ESL contexts. In Hamp-Lyons, L., editor, *Assessing second language writing in academic contexts*. Norwood: Ablex, 241–76.

Hamp-Lyons, L. 1995: Rating nonnative writing: the trouble with holistic scoring. *TESOL Quarterly* 29, 759–62.

Hamp-Lyons, L. and **Henning, G.** 1991: Communicative writing profiles: an investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning* 41, 337–73.

Hinds, J. 1987: Reader vs. writer responsibility: a new typology. In Connor, V. and Kaplan, R.B., editors, *Writing across languages: analysis of L2 text*. Reading, MA: Addison-Wesley, 141–52.

Hirose, K. and **Kobayashi, H.** 1991: Cooperative small group discussion. *JALT Journal 13*, 57–72.

Hirose, K. and **Sasaki, M.** 1994: Explanatory variables for Japanese students' expository writing in English: an exploratory study. *Journal of Second Language Writing* 3, 203–29.

Huot, B. 1990: The literature of direct writing assessment: major concerns and prevailing trends. *Review of Education Research* 60, 237–63.

Jacobs, H.L., Zinkgraf, S.A., Wormuth, D.R., Hartfiel, V.F. and **Hughey, J.B.** 1981: *Testing ESL composition: a practical approach*. Rowley, MA: Newbury House.

Kinoshita, K. 1981: *Rikakei no Sakubun Gijutu* [Writing techniques for science writing]. Tokyo: Chuo Koronsha.

Kinoshita, K. 1990: *Repooto no Kumitatekata* [How to write a report]. Tokyo: Chikuma Shobou.

Kitagawa, M. and **Kitagawa, C.** 1987: *Making connections with writing: an expressive writing model in Japanese schools*. Portsmouth, NH: Heinemann.

Kokugo Kyoiku Kenkyusho. 1988: *Kokugo Kyouiku Kenkyuu Daijiten* [Dictionary of Japanese language education]. Tokyo: Meiji Tosho.

Komatsu, Z. 1976: *Kokugo no Jugyou Soshikiron* [Systematization of Japanese language curriculum]. Tokyo: Ikkosha.

Messick, S. 1993: Validity. In Linn, R. L., editor, *Educational measurement*. 3rd edition. New York: American Council on Education and Oryx Press, 13–103.

Ministry of Education, Science and Culture, Government of Japan. 1989: *Course of study for upper secondary school in Japan*. Tokyo: Printing Bureau, Ministry of Finance.

Oouchi, Z. 1995: Seikatsu Tsuzurikata to Sakubunkyoiku Ronsou [Controversies between the life-experience writing school and the expression school]. *Kyouiku Kagaku Kokugo Kyouiku* [Educational Science of Teaching Japanese], 48–51.

Pennington, M.C. and **So, S.** 1993: Comparing writing process and product across two languages: A study of 6 Singaporean university student writers. *Journal of Second Language Writing* 2, 41–63.

Raimes, A. 1990: The TOEFL test of written English: causes for concern. *TESOL Quarterly* 24, 427–42.

Sasaki, M. 1996: *Second language proficiency, foreign language aptitude, and intelligence: quantitative and qualitative analyses*. New York: Peter Lang.

Sasaki, M. and **Hirose, K.** 1996: Explanatory variables for EFL students' expository writing. *Language Learning* 46, 137–74.

Shiojiri, Y. 1978: *Kokugoka no Totasudo Hyoka Nyuumon* [Introduction to formative assessment in Japanese]. Tokyo: Meijitosho.

Tabachnick, B.G. and **Fidell, L.S.** 1989: *Using multivariate statistics*. 2nd edition. New York: Harper & Row.

Ueyama, T and **Morita, N.** 1990: Setsumeiteki Bunsho no Shido [Teaching expository writing in Japanese as L1]. In Otsuki, K., editor, *Kokugo Kyoikugaku* [Japanese language education]. Tokyo: Fukumura Shuppan, 103–30.

Appendix 1 Questionnaire on Japanese composition evaluation*

Please provide the following information. You do not have to write your name.

In which prefecture is your school located?: (   ) Prefecture

Sex: Male Female Age:
Years of teaching experience:

Suppose you evaluate expository compositions (e.g., the topics of those compositions are such as “Should married women work outside their home?,” “Are you for or against those who do not have a steady job?”). How much importance do you put on the following 35 items? Please circle the number which best fits your judgment from 1 (not at all important) to 5 (very important).

		Not at all	*Little*		*Fairly*	*Very*
	(EXPRESSION)					
1.	Is the handwriting easy to read?	1	2	3	4	5
2.	Is the notation (e.g., letters, punctuation marks, *kana* orthography) correct?	1	2	3	4	5
3.	Is word usage correct?	1	2	3	4	5
4.	Is vocabulary rich?	1	2	3	4	5
5.	Are sentences well-formed?	1	2	3	4	5
6.	Are sentences sufficiently short?	1	2	3	4	5
7.	Is ‘neutral style’ distinguished from ‘polite style’ (with *desu-masu* verb forms)?	1	2	3	4	5
8.	Are there any grammatical mistakes?	1	2	3	4	5
9.	Are sentences adequately connected with appropriate use of conjunctions and demonstrative words?	1	2	3	4	5
10.	Are sentences adequately connected in terms of meaning and logic?	1	2	3	4	5
11.	Are sentences sufficiently concise?	1	2	3	4	5
12.	Are sentences unambiguous? (Can they be interpreted in more than one way?)	1	2	3	4	5
13.	Are various rhetorical expressions used appropriately?	1	2	3	4	5

		Not at all	*Little*		*Fairly*	*Very*
	(ORGANIZATION)					
14.	Are paragraphs appropriately formed?	1	2	3	4	5
15.	Are all paragraphs logically connected?	1	2	3	4	5
16.	Is the main point written at the beginning of a paragraph?	1	2	3	4	5
17.	Is there a concluding paragraph?	1	2	3	4	5
18.	Do paragraphs follow a general organizational pattern such as “introduction–body–conclusion” or “*ki-shoo-ten-ketsu*”?**	1	2	3	4	5
	(CONTENT)					
19.	Are facts and opinions differentiated?	1	2	3	4	5
20.	Are facts and examples provided based on the writer’s experience?	1	2	3	4	5
21.	Are facts and examples provided based on the writer’s concrete experience and knowledge?	1	2	3	4	5
22.	Is the theme clear?	1	2	3	4	5
23.	Is the theme supported by sufficient factual information?	1	2	3	4	5
24.	Does the writer take a clear position “for” or “against” the given opinion?	1	2	3	4	5
	(APPEAL TO THE READERS)					
25.	Are paragraphs ordered so that it is easy for the reader to follow?	1	2	3	4	5
26.	Are expressions and notation easy for the reader to understand? Are any expressions too complicated?	1	2	3	4	5

		Not at all	*Little*		*Fairly*	*Very*
27.	Are given facts and reasons easy for the reader to understand?	1	2	3	4	5
28.	Is there any appealing content provided?	1	2	3	4	5
29.	Is there any surprising/novel content provided?	1	2	3	4	5
	(SOCIAL AWARENESS)					
30.	Does the writer demonstrate objective awareness of him/herself?	1	2	3	4	5
31.	Does the writer attempt to look at him/herself in a new light?	1	2	3	4	5
32.	Does the writer demonstrate objective awareness of social phenomena?	1	2	3	4	5
33.	Does the writer attempt to look at social phenomena in a new light?	1	2	3	4	5
34.	Does the writer demonstrate objective awareness of the relationship between the society and him/herself?	1	2	3	4	5
35.	Does the writer attempt to look at the relationship between society and him/herself in a new light?	1	2	3	4	5

Notes: *The original version was written in Japanese; **Japanese teachers are familiar with the traditional Japanese writing pattern called ‘ki-shooten-ketsu’. According to this pattern, the writer first introduces the topic in ‘ki’, and develops the topic in ‘shoo’; and makes an abrupt transition in ‘ten’; and finally concludes the topic in ‘ketsu’.

Appendix 2a Rating scale for Japanese L1 expository writing

国語説明文評価表

学生の名前 日付

評点	基準		
＊主題の明確性	１０～９	大変良い	○主題が明確である。主題を根拠づけるのに十分な事実が書かれている。事実と意見とを区別して書いている。
	８～６	良い	○主題がある程度明確である。主題のための根拠・事実がある程度書かれている。
	５～３	あまり良くない	○主題があまり明確でない。主題のための根拠・事実に乏しい。
	２～１	良くない	○主題が全く明確でない。
＊読者に対する説得性	１０～９	大変良い	○具体的な根拠・事例が用いられており、説得力がある。読み手が共鳴する内容を持っている。
	８～６	良い	○具体的根拠・事例が用いられており、ある程度説得力がある。読み手が共鳴するような内容が、ある程度書かれている。
	５～３	あまり良くない	○具体的な根拠・事例が少なく、あまり説得力がない。読み手に訴えるような内容に乏しい。
	２～１	良くない	○具体的な根拠・事例がほとんど用いられておらず、読み手に訴えるような内容が無い。
＊表現	１０～９	大変良い	○文が首尾一貫していて、文と文が、適切につながっている。
	８～６	良い	○それぞれの文は首尾一貫しているが、文と文が適切につながっていない箇所がある。
	５～３	あまり良くない	○文が首尾一貫していないことがあり、また、文と文のつながりが、不適切な箇所が多い。
	２～１	良くない	○文が首尾一貫していず、文と文のつながりが、非常に不適切である。
＊構成	１０～９	大変良い	○段落相互の意味、論理関係が適切で、段落のつながりが、読み手にわかりやすい順序になっている。
	８～６	良い	○段落相互の意味、論理関係が、ある程度適切で、段落のつながりが、ある程度読み手にわかりやすい順序になっている。
	５～３	あまり良くない	○段落相互の意味、論理関係があまり適切でなく、段落のつながりが読み手にわかりづらい。
	２～１	良くない	○段落相互の意味、論理関係が不明で、段落のつながりがわからない。
＊形式的言語知識	１０～９	大変良い	○適切な表記（文字、句読点、送り仮名、漢字使用等）に従っている。正しい意味で語が用いられている。文法の間違いがない。
	８～６	良い	○表記、用語、文法に、ときどき不適切な箇所がある。
	５～３	あまり良くない	○表記、用語、文法に、しばしば不適切な箇所がある。
	２～１	良くない	○表記、用語、文法が不適切である。
＊書き手の対象認識	１０～９	大変良い	○書き手が、自己、社会の事象、及び自己と社会の関係を認識しようとしている。
	８～６	良い	○書き手が、自己、社会の事象、及び自己と社会の関係を認識しようとしているのが、ある程度うかがえる。
	５～３	あまり良くない	○書き手が、自己、社会の事象、及び自己と社会の関係を認識しようとしているのが、あまりうかがえない。
	２～１	良くない	○書き手が、自己、社会の事象、及び自己と社会の関係を全く認識しようとしていない。
合計点		／６０点	

Appendix 2b Rating scale for Japanese L1 expository writing (translation)

Name: Date:

SCORE	CRITERIA		
Clarity of the theme	10–9	very good	● Theme is clear. ● Provides sufficient facts to support the theme. ● Differentiates facts from opinions.
	8–6	good	● Theme is somewhat clear. ● Provides some facts and reasons to support the theme.
	5–3	fair	● Theme is not so clear. ● Provides few facts and reasons to support the theme.
	2–1	poor	● Theme is not clear at all.
Appeal to the readers	10–9	very good	● Provides concrete and convincing reasons and facts. ● Very appealing to the reader.
	8–6	good	● Provides somewhat concrete and convincing reasons and facts. ● Appealing to the reader.
	5–3	fair	● Provides a few concrete and convincing reasons and facts. ● Not so appealing to the reader.
	2–1	poor	● Provides few concrete and convincing reasons and facts. ● Not appealing to the reader.
Expression	10–9	very good	● All sentences are consistently structured and adequately connected.
	8–6	good	● All sentences are consistently structured, but some sentences are inadequately connected.
	5–3	fair	● Not all sentences are consistently structured, and many sentences are inadequately connected.
	2–1	poor	● Sentences are inconsistently structured and are inadequately connected.
Organization	10–9	very good	● All paragraphs are logically connected, and easy to follow.
	8–6	good	● All paragraphs are somewhat logically connected, and not difficult to follow.
	5–3	fair	● Paragraphs are not logically connected, and difficult to follow.
	2–1	poor	● All paragraphs are not logically connected at all, and impossible to follow.
Knowledge of language forms	10–9	very good	● Follows appropriate notation (spelling, punctuation, correct use of Chinese characters, etc.). ● Demonstrates mastery of correct word usage and grammar.
	8–6	good	● Sometimes makes errors in notation, word usage, and grammar.
	5–3	fair	● Often makes mistakes in notation, word usage, and grammar.
	2–1	poor	● Demonstrates no mastery of notation, word usage, and grammar.

Continued

Appendix 2b Continued

SCORE	CRITERIA		
Social awareness	10–9	very good	● Demonstrates full awareness of oneself, social phenomena, and the relationship between oneself and society.
	8–6	good	● Demonstrates some awareness of oneself, social phenomena, and the relationship between oneself and society.
	5–3	fair	● Demonstrates little awareness of oneself, social phenomena, and the relationship between oneself and society.
	2–1	poor	● Demonstrates no awareness of oneself, social phenomena, and the relationship between oneself and society.

Total score /60 points