Pan-Pacific Association of Applied Linguistics 13(2), 77-103

The Development and Implementation of Task-based Writing Performance Assessment for Japanese Learners of English

Yoshihito Sugita
Yamanashi Prefectural University

Sugita, Y. (2009). The development and implementation of task-based writing performance assessment for Japanese learners of English. *Journal of Pan-Pacific Association of Applied Linguistics*, 13(2), 77-103.

The main purpose of this research is 1) to establish a framework for the test development and the constructs of writing performance test, 2) to implement a developed writing performance assessment, and 3) to examine the degree of reliability and validity of the assessment tasks and rating scales. Construct-based processing approach to testing resulted in a comprehensive framework for our test development. Accuracy and communicability were defined as constructs, and the test development proceeded according to the three stages. The test was conducted as an examination into the assessment tasks and rating scales, and the analyses were done using FACETS. The results showed that 1) the difficulty of the two tasks and the impressionistic scoring were considered equivalent, which provided reasonable fit to the Rasch model, 2) the equivalence of task difficulty may indicate that task development based on construct-based processing approach could be reliable and valid to estimate students' writing ability, and 3) the rating scales associated with the five rating categories and their specific written samples were shown to be mostly comprehensible and usable by raters, and demonstrated acceptable fit. However, there is still room for argument about the reliability and validity of assessment tasks and rating scales.

Key Words: writing performance, task-based assessment, FACETS, reliability, validity

1 Introduction

In Japan English language has been traditionally taught with a focus on accuracy, and indirect measurement is widely used in the field of assessment. There seems to have been a paradigm shift from accuracy-oriented to fluency-oriented writing instruction, but no significant changes have occurred in assessment of writing. Judging from the present state of teaching and assessing writing in Japan, it would be meaningful to develop scoring procedures for writing performance assessment in place of traditional indirect tests of writing. This study is motivated by such an urgent need for improved

assessment of writing, which is conducted in order to develop a task-based writing test for Japanese learners of English.

2 Development of Task-based Writing Test (TBWT)

2.1 Construct-based processing approach to testing

As Bachman and Palmer mentioned (1996), the primary purpose of a language test is to make inferences about language ability. The ability that we want to test is defined as a construct, and describing the construct is one of the most fundamental concerns in test development. When assessing writing, it is therefore necessary to address the issue of how much importance we place on the ability of our students to write.

Skehan (1998) claimed that the processing perspective is relevant to the way directly explaining underlying abilities to performance and how we conceive of models of language ability. In this view, he defines "ability for use" as a construct, which rationalizes the use of tasks as a central unit within a testing context and in developing a performance test. According to Skehan, such a task-based approach to testing would be "to assume that there is a scale of difficulty and that students with greater levels of underlying ability will then be able to more successfully complete tasks which come higher on such a scale of difficulty" (p.174). In this assumption, we find that task difficulty is a major determinant of test performance. Task-based approaches, therefore, need to focus on task difficulty as a precondition for using tasks-as-tests, and methods of evaluating task-based performance. Bachman (2002), however, claimed that task difficulty can be found with the various components in a performance task and with the interactions among them, and thus task difficulty is not a separate factor and is no longer assumed to be a major determinant of test performance. Therefore, he emphasized that the task-based approach has to consider not only performances on tasks, but also abilities to be assessed. In this way, Bachman argues that the view of construct-based approach to testing is also necessary for test development, and mentions that the most important thing is to integrate tasks and construct in the design and development of a particular assessment.

Here, we notice that there is considerable validity in the integration of construct-based task development and task implementation based on the operation of the processing factors and the influences of the processing conditions. In other words, when we develop assessment tasks, it is reasonable to suppose that we should design the task on the basis of construct definition and processing perspectives. Thus, the so-called *construct-based processing approach to testing* results in a comprehensive framework for our test development. The characteristic features of this approach are: 1) it must consider both

constructs and tasks in developing performance assessment (Bachman, 2002); 2) procedures for design, development and use of language tests must incorporate both a specification of the assessment tasks to be included and definitions of the abilities to be assessed (Alderson et al., 1995; Bachman & Palmer, 1996; Brown, 1996); 3) tasks should be conceptualized as sets of characteristics (Bachman, 2002), and task characteristics should be designed to consider performance on tasks in terms of the operation of the processing factors and the influences of the processing conditions (Skehan, 1998); and 4) the processing factors that affect performance such as communicative stress should be utilized in order to control processing conditions in which it involves the interaction of test-taker attributes.

2.2 Construct definition

The constructs of our task-based writing test developed for this study are assumed to be *accuracy* and *communicability*. Both constructs are derived from the Bachman and Palmer framework (1996) and the Skehan's processing perspective on testing (1998). As shown in Figure 1, *accuracy* shares the rule-based system in terms of the processing perspectives, and has a deep connection with organizational knowledge which consists of grammatical and textual knowledge. Grammatical knowledge "is involved in producing or comprehending formally accurate utterances or sentences," and textual knowledge "is involved in producing or comprehending texts that consists of two or more utterances or sentences" (Bachman & Palmer, 1996, p.68). Based on these two areas of organizational knowledge, it is proposed here that the construct *accuracy* specialized for writing would be comprised of organizational skills and linguistic accuracy. Specifically, organizational skills can be defined as the ability to organize logical structure which enables the content to be accurately acquired, and linguistic accuracy concerns errors of vocabulary, spelling, punctuation or grammar (Sugita, 2008).

Figure 1. The construct structure of accuracy

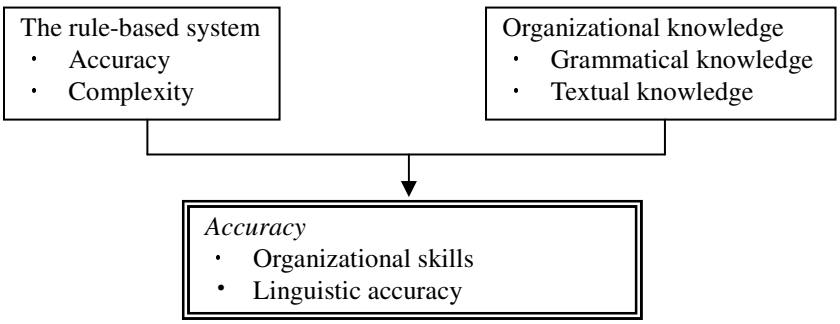
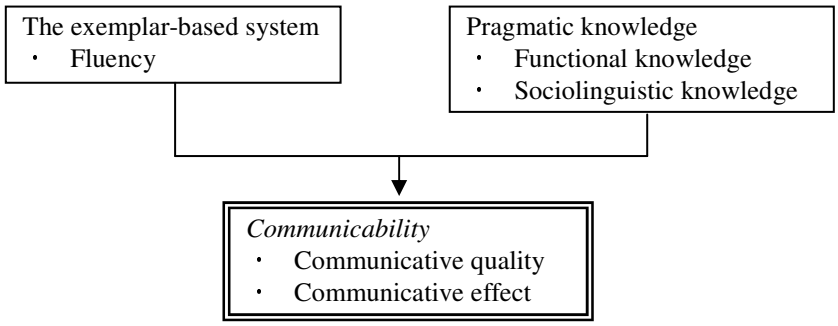


Figure 2 indicates the construct structure of *communicability*. We realize that the construct shares the exemplar-based system in terms of the processing perspectives, and its basis is pragmatic knowledge which consists of functional and sociolinguistic knowledge. Functional knowledge "enables us to interpret relationships between utterances or sentences and texts and the intentions of language users" (Bachman & Palmer, 1996, p.69). Sociolinguistic knowledge enables us to create or interpret language that is appropriate to a particular language use setting (p.70). Based on these definitions and the processing perspectives, the term *communicability* is defined as fluency specialized for writing, which is comprised of communicative quality and effect. Communicative quality refers to the ability to communicate without causing the reader any difficulty, and communicative effect concerns the quantity of ideas necessary to develop the response as well as the relevance of the content to the proposed task (Sugita, 2008).

Figure 2. The construct structure of communicability



2.3 Procedures for developing the TBWT

In terms of construct-based processing approach, the test development proceeded according to the following three stages:

Stage 1: Designing and characterizing writing tasks

With regard to processing perspectives (Skehan, 1998), content-based support and form-focused stakes are necessary for accuracy tasks. An elicitation task (writing a letter) was chosen, and specific topics of self-introduction were given in the task. A situation is supposed in which the student is going to stay with a host family in Britain, and is suggested to write a letter, so that students can focus on writing accuracy. On the contrary, communicability tasks need form-oriented support and meaning-focused stakes in order to write with a focus on

meaning. A discussion task was designed because it encourages students to write their opinions or ideas about the topic, and it lays emphasis on meaning-focused response (see the specifications in Appendix A).

According to Bachman and Palmer (1996), characteristics of the input and the expected response in a test task are closely concerned with the operation of the processing factors and influences of the processing conditions for task implementation. In view of the construct-based processing approach to testing, the TBWT needs to develop such task characteristics in order to adjust students to performance conditions in which they allocate attention in appropriate ways. Specifically, characteristics in accuracy tasks require students to write a 100-120 word letter in adequate time in order that the rule-based system can be accessed, and characteristics in communicability tasks encourage students to write as many answers to a discussion topic as possible in very limited time in order that an exemplar-based system will be appropriate.

Stage 2: Reviewing existing scoring procedures for assessing writing

Existing scoring procedures for assessing writing were considered in order to explore what types of procedures are more suitable to construct rating scales. Eventually, the TBWT is as construct-relevant as multiple trait scoring, and its procedure is similar to primary trait scoring in that scoring criteria are developed for each elicitation task. In such a combined procedure, the two assessment tasks and their criteria exist independently, and thus raters are required to make only one decision for every script as conducted in holistic scoring.

Stage 3: Drafting rating scales

The underlying competences served as a useful basis when developing rating scales for accuracy and communicability. The descriptors of the marking categories in each scale were collected from the existing writing assessment such as the TOEFL Test of Written English (TWE) and Cambridge First Certificate in English (FCE). By conforming one construct closely to the definition of its rating scale, it is fair to say that raters would use the scale appropriately and consistently, ensuring the reliability and validity of assessing writing. According to Alderson *et al*. (1995), raters should understand the principles behind the particular rating scales they must work with, and be able to interpret their descriptors consistently. Therefore, the rating scales are comprised of clearer descriptions of each construct and of 5-point Likert scales (Appendix B). The descriptors of each category are also provided with the selected written samples as an explanatory part of the scale in order that busy school teachers with limited training on writing

performance assessment can understand the descriptors and work with them consistently (Appendix C).

3 The Study

3.1 Purposes and research questions

In order to examine the degree of reliability and validity of the task-based writing performance test, the following are focused on: raters' severity, interactions with writers' abilities and task difficulties, the reliability of elicitation tasks and rating scales, and the measure's validity. The specific research questions are as follows:

1) Is student ability effectively measured?
2) Are teacher-raters equally severe?
3) How much do tasks that are designed to be equivalent actually differ in difficulty?
4) How well do scales conform to expectations about their use? Do raters use all parts of them, and use them consistently?
5) Do individual raters score a particular group of subjects more harshly or more leniently? If so, what are the sub-patterns of ratings in terms of rater-subject interaction for each rater?
6) Do the raters score particular tasks more harshly or more leniently than others? If so, what are the sub-patterns of ratings in terms of rater-task interaction for each rater?
7) To what extent, statistically, is the task-based writing test a reliable and valid measure?

3.2 Test participants and materials

The data for this study were 40 scripts (20 scripts for each of two tasks) collected from 20 undergraduate students (6 males and 14 females) who took an English teaching methodology course in the first semester of 2008. The subjects were 14 second-year and 6 third-year students from the faculty of global policy management and communications. All of the subjects were native speakers of Japanese with an intermediate level of English language proficiency.

The TBWT was conducted in the computer-assisted learning room. The time limits for Task 1 and Task 2 were 20 minutes and 10 minutes, respectively. After finishing the tasks, each student was required to submit an essay using a web-based essay evaluation service, *Criterion*. They had to finish writing the essay within 30 minutes. The prompt was as follows: People attend college or university for many different reasons (for example, new experiences, career preparation, increased knowledge). Why do you

think people attend college or university? Use specific reasons and examples to support your answer.

Criterion provides immediate score reporting and diagnostic feedback on students' essays. The system is comprised of an E-rater scoring engine and a Critique writing analysis tool. E-rater assigns a holistic score to an essay on a 6-point scale by comparing its linguistic features to those of the human-scored essays stored in the system's database. Critique detects errors in grammar, usage, and mechanics and identifies undesirable style and essay-based discourse elements (Burstein, Chodorow, & Leacock, 2003). Burstein *et al*. (2003) have found that there is usually 97% agreement on holistic scores between E-raters and human raters, which is as high as the inter-rater reliability of two human raters. This validation study indicates that *Criterion* has a high internal consistency as a writing performance test. The scoring can solve the subjectivity problem inherent in writing assessment, and I expect that the construct validity of the two tasks and impressionistic scoring can be discussed by examining their scores and those with a high reliability provided by *Criterion*.

3.3 Scoring materials and procedure

Some previous studies (ex. Shohamy et al., 1992; Weigle, 1994) implied that a thorough understanding of the ability being measured by the test might be a central aspect of the training process if the raters behave consistently. This view of the function of training addresses the concern that the scoring guide which gives raters a shared understanding of the construct of writing ability as defined by the test writers may effectively reduce the differences or biases caused by variation among raters. For this purpose, the TBWT scoring guide was edited for this testing. The first section is the background of the TBWT. The second section is the explanation of assessment tasks. The third section is the implementation method of the testing. The fourth section is comprised of the rating scales and written samples accompanied by detailed commentary on each sample at five levels, 1-5 (see Appendix B & C).

Each of the forty scripts was scored by five raters, who were all experienced Japanese high school teachers of English. They were all native speakers of Japanese, and they shared similar backgrounds in terms of qualifications of ten or more years of teaching experience. They displayed acceptable levels of consistency with themselves in the pre-testing conducted in January, 2008. Both scripts and scoring guidelines were given to the raters by mail at the end of July, 2008. Each of the five raters rated the entire set of forty scripts and sent them back by the end of August, 2008. They were instructed to rate the 20 scripts of Task 1 first, and then to rate the 20 scripts of Task 2. Finally, they were asked to rate each of the participants' writing proficiency based on the total impression at five levels, 1-5.

3.4 Data analysis

Table 1, 2 and 3 show the descriptive statistics for the scores of the two test tasks and the impressionistic scoring. Table 4 summarizes the inter-rater correlation coefficients for the different scoring. Since the average of the coefficients for each scoring is relatively high (0.76, 0.83, 0.81), the five raters appear to have demonstrated acceptable reliability.

Table 1. Descriptive Statistics of Scoring Task 1

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Mean	3.20	3.10	3.50	2.85	2.40
SD	0.87	1.30	1.00	1.19	0.91
Minimum	2.0	1.0	1.0	1.0	1.0
Maximum	5.0	5.0	5.0	5.0	5.0

Table 2. Descriptive Statistics of Scoring Task 2

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Mean	2.95	3.10	3.05	2.65	2.85
SD	0.97	1.37	1.20	1.19	0.96
Minimum	1.0	1.0	1.0	1.0	1.0
Maximum	5.0	5.0	5.0	5.0	5.0

Table 3. Descriptive Statistics of Impressionistic Scoring

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Mean	3.10	3.20	3.25	2.85	2.50
SD	0.88	1.20	1.13	1.23	1.57
Minimum	2.0	1.0	1.0	1.0	1.0
Maximum	5.0	5.0	5.0	5.0	5.0

Table 4. Inter-rater Correlation Coefficients between Pairs of Raters

R	1/2	1/3	1/4	1/5	2/3	2/4	2/5	3/4	3/5	4/5	Av
T1	.82	.76	.79	.52	.74	.91	.72	.79	.72	.83	.76
T2	.78	.77	.84	.79	.84	.93	.84	.91	.78	.82	.83
IS	.77	.76	.78	.60	.87	.89	.88	.88	.83	.88	.81

Note. R=rater; T1=task 1; T2=task 2; IS=impressionistic scoring; Av.=average

Table 5 reports results for each test task, the impressionistic scoring and the scores of *Criterion*, including its mean and standard deviation. The mean scores for all variables are very close, ranging from 2.92 to 3.01. The alpha coefficients for the test tasks and the impressionistic scoring were calculated. Using Davies' cut-off (.90) as an acceptable level of internal consistency on a high-stakes test, each Cronbach's α would meet the point: .9386, .9582 and .9570 for Task 1, Task 2 and impressionistic scoring, respectively.

Table 5. Descriptive Statistics of the Different Scoring

	Task 1	Task 2	Impression	*Criterion* TWE
N	100	100	100	20
Mean	3.01	2.92	2.98	2.40
SD	1.12	1.16	1.14	0.94
Minimum	1.0	1.0	1.0	1.0
Maximum	5.0	5.0	5.0	4.0

The correlation coefficients between the scores provide a preliminary estimate of the parallel-form reliability of each test task. As seen in Table 6, the correlation coefficients between each task and the impressionistic score fall in a range of .778 to .910, which are all significant at the 0.01 level. The correlation between the two test tasks (.778) is, however, slightly lower than the established estimate of reliability (.80). Table 6 also shows that the two tasks and impressionistic scoring correlate positively with the scores of *Criterion* ($p<.01$). The highest correlation is between the *Criterion* score and Task 2 ($r=.708$), followed by that between the *Criterion* score and Impression ($r=.703$) and finally, between the *Criterion* score and Task 1 ($r=.621$).

Table 6. Pearson Correlation Coefficients

	Task 1	Task 2	Impression
Task 2	.778**		
Impression	.910**	.903**	
Criterion	.621**	.708**	.703**

Note. **all correlations significant at 0.01 level

There is a possibility that the test data can be influenced by errors of measurement resulting from variation in rater harshness and test tasks, as well as by the nature of the rating scale used and by the range of ability of the subjects who are being assessed. Therefore, it was necessary to use statistical models which take into account all of the factors that might affect a student's final score.

The analyses for the present study were done using FACETS version 3.63 (Linacre, 2008). To examine the measurement characteristics of this testing, the data was specified as having three facets, namely, the ability of the subjects, the difficulty of tasks and the severity of raters. The partial-credit model was chosen because the scoring criteria for the rating scales were qualitatively different.

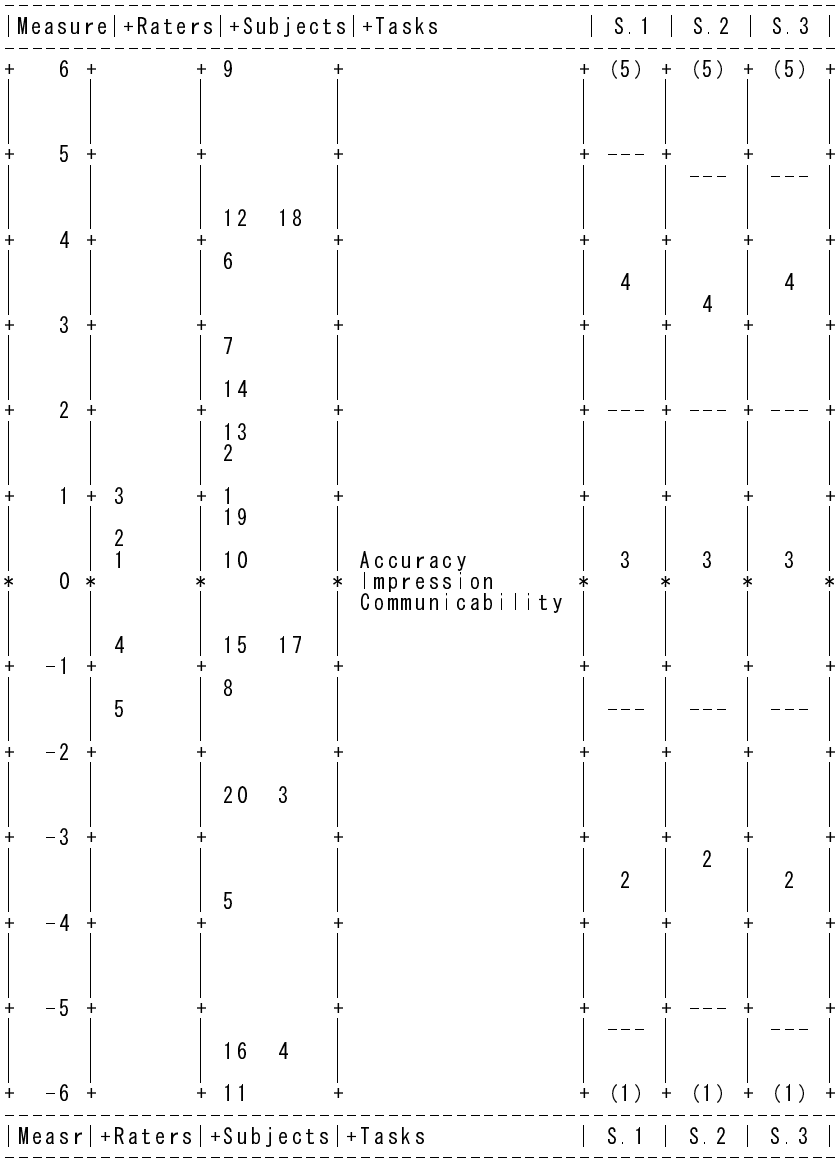
4 Results

4.1 FACETS summary

Figure 3 shows a summary of all facets and their elements. They are positioned on a common logit scale, which appears as "measure" in the first

column. The second column shows the severity variation among raters. The most severe rater (ID: 3) is at the top, and the least severe rater (ID: 5) is at the bottom. The third column shows the ability variation among the 20 subjects. The subjects are ranked with high ability at the top (ID: 9) and low ability at the bottom (ID: 11).

Figure 3. FACETS summary



The fourth column shows the difficulty variation among tasks. The most severely scored task (Accuracy) is at the top and the least severely scored task (Communicability) is at the bottom. The last three columns graphically describe the three rating scales. Each of the two tasks and the impressionistic scoring has their own scale. The most likely scale score for each ability level is shown.

4.2 FACETS analysis

1) *Is student ability effectively measured?*

As shown in Figure 1, subject ability estimates range from a high of 5.94 logits to a low of −5.92 logits, indicating a spread of 12 logits in terms of students' ability. Subject separation value was 6.85, meaning that populations like the students in this study can be spread into about seven levels. The reliability index was .98, which demonstrates the possibility to achieve reliable ability scores.

2) *Are teacher-raters equally severe?*

Table 7. FACETS Analysis of Rater Characteristics

	Fair-M average	Severity (logits)	Error	Infit (mean square)
Rater 1	3.03	.35	.24	.97
Rater 2	3.07	.52	.24	.93
Rater 3	3.19	.97	.24	.78
Rater 4	2.75	-.69	.24	.65
Rater 5	2.54	-1.41	.25	1.16
Mean	2.92	-.05	.24	.90
SD	.23	.87	.00	.17

Note. Reliability of separation index=.92; fixed (all same) chi-square: 62.8, df:4; significance: p=.00

Table 7 provides information on the characteristics of raters. From the left, each column shows rater IDs, fair average scores, rater severity, error and fit mean square values. The second column indicates that the severity span between the most severe rater and the most lenient rater was 2.38 and the difference, based on fair average scores in the first column, is 0.65 of one grade in the scale. The reliability of the separation index (which indicates the likelihood to which raters consistently differ from one another in overall severity) was high (.92). The chi-square of 62.8 with 4 df was significant at $p<.00$ and, therefore, the null hypothesis that all raters were equally severe must be rejected. There was a significant difference in severity among raters. On the other hand, the Infit Mean Square column indicates that no raters were identified as misfitting: fit values for all raters were within the range of two

standard deviations around the mean (0.90±[0.17×2]). In other words, all raters behaved consistently in the scoring.

3) *How much do tasks that are designed to be equivalent actually differ in difficulty?*

The analysis of the two test tasks and impressionistic scoring in Table 8 shows that no significant variation in difficulty exists among them. Raters are considered to be self-consistent in scoring, and the tasks do not appear to separate the subjects to a significant degree meaning that the difficulty of the two tasks and the total impression of the tasks can be considered equivalent. An estimate of the item discrimination was computed according to the "Generalized Partial Credit Model" approach. 1.0 is the expected value, but discriminations in the range 0.5 to 1.5 provide a reasonable fit with the Rasch model (Linacre, 2007, p.132).

Table 8. Descriptive statistics on the different scoring

	Difficulty (logits)	Error	Infit (mean square)	Estimate of Discrimination
Task 1	.13	.19	1.10	.90
Task 2	-.18	.19	.92	1.05
Impression	.05	.19	.68	1.37
Mean	.00	.19	.90	
SD	.13	.00	.17	

Note. Reliability of separation index=.00; fixed (all same) chi-square: 1.5, df:2; significance: p=.47

4) *How well do scales conform to expectations about their use? Do raters use all parts of them, and use them consistently?*

Linacre (2002) has proposed the following guidelines for a rating scale: (1) average category measures should advance monotonically according to the category, (2) outfit mean-squares should be less than 2.0, and (3) the step difficulty of each scale should advance by at least 1.4 logits and by no more than 5.0 logits.

Table 9. Rating scale statistics for Accuracy

Category Score	Average Measure	Outfit (mean square)	Step Difficulty
1	-5.77	.9	
2	-2.69	1.1	-5.28
3	.26	1.4	-1.54
4	2.87	.8	1.96
5	4.95	1.0	4.86

Figure 4. Probability curves for accuracy

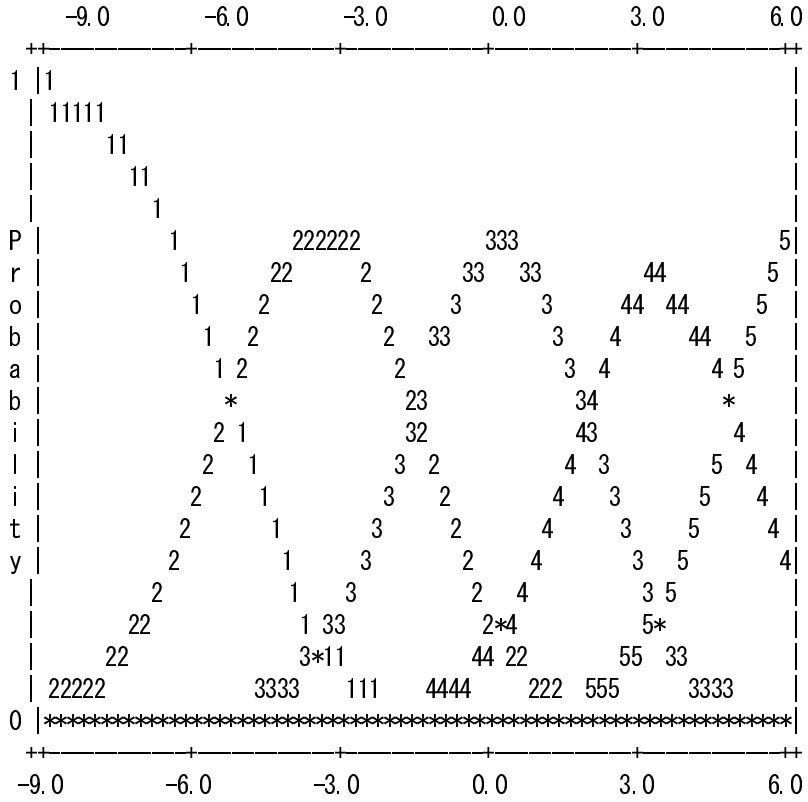


Table 9 shows the rating scale statistics for accuracy. Since higher category scores are intended to reflect higher measures, the average category measures are expected to rise. All outfit mean-squares are less than 2.0, meaning that each of the five categories has expected randomness in choosing categories. All increases in step difficulty fall within 1.4 and 5.0, which does meet (3). The scale structure probability curves are shown in Figure 4. Starting from the left, category 1 is most likely to be observed for low-measure scripts. Then as script measures increase, the probability of observing category 2 increases. With increasing measure, category 3 becomes most probable, then 4, and finally 5. According to Linacre (1999), if the modeled category probability curves depict a succession of “hills”, the step difficulties successively increase with category scores, meaning that each category in turn is most likely to be chosen. Tyndall and Kenyon (1995) mentioned that the obvious peaks and divisions between the categories indicate that the scales conform to the expectations regarding their use. In Figure 2, the curves are like the expected succession of hills and obvious hill

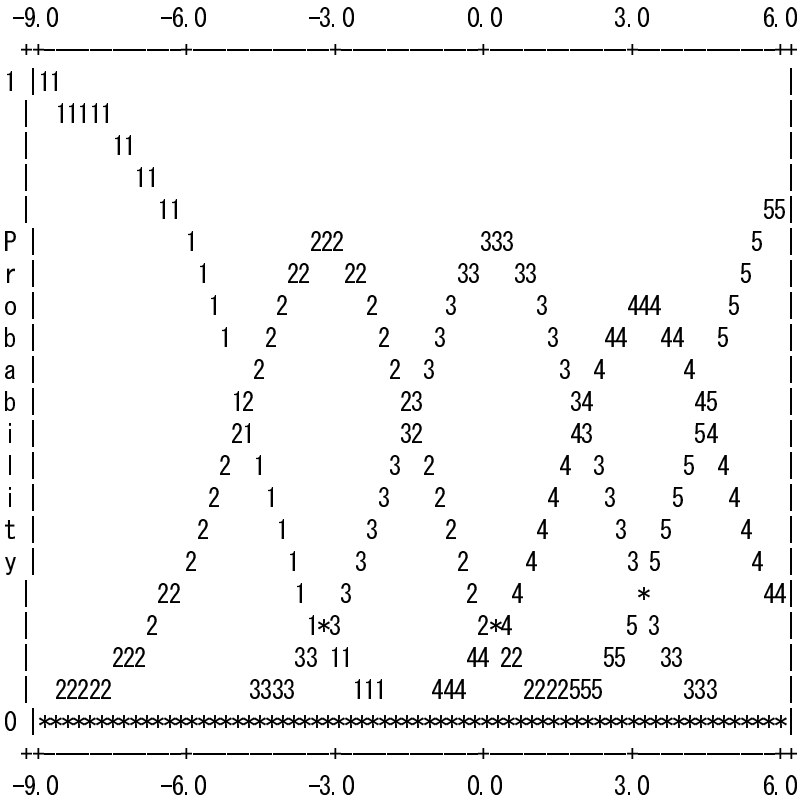
tops are observed, which imply that the rating scale can be decomposed into five categories.

Table 10. Rating Scale Statistics for Communicability

Category Score	Average Measure	Outfit (mean square)	Step Difficulty
1	-5.35	1.2	
2	-3.04	1.6	-4.97
3	.26	.9	-1.49
4	2.74	.6	1.98
5	4.97	.7	4.49

Table 10 shows the rating scale statistics for communicability. All outfit mean-squares are less than 2.0, which meet (2). All step difficulty increases fall within 1.4 and 5.0, which does meet (3).

Figure 5. Probability curves for communicability



In Figure 5, the plot of the rating scale probability curves which depicts a range of hills indicates that the step difficulties become successively more positive as the rating scale increases. The obvious peaks and the division between the scales also indicate that the scales work as intended.

Table 11. Rating Scale Statistics for Impression

Category Score	Average Measure	Outfit (mean square)	Step Difficulty
1	-5.87	.7	
2	-3.19	.5	-5.13
3	.55	.7	-1.65
4	2.96	.7	2.27
5	5.05	.8	4.51

Figure 6. Probability curves for impression

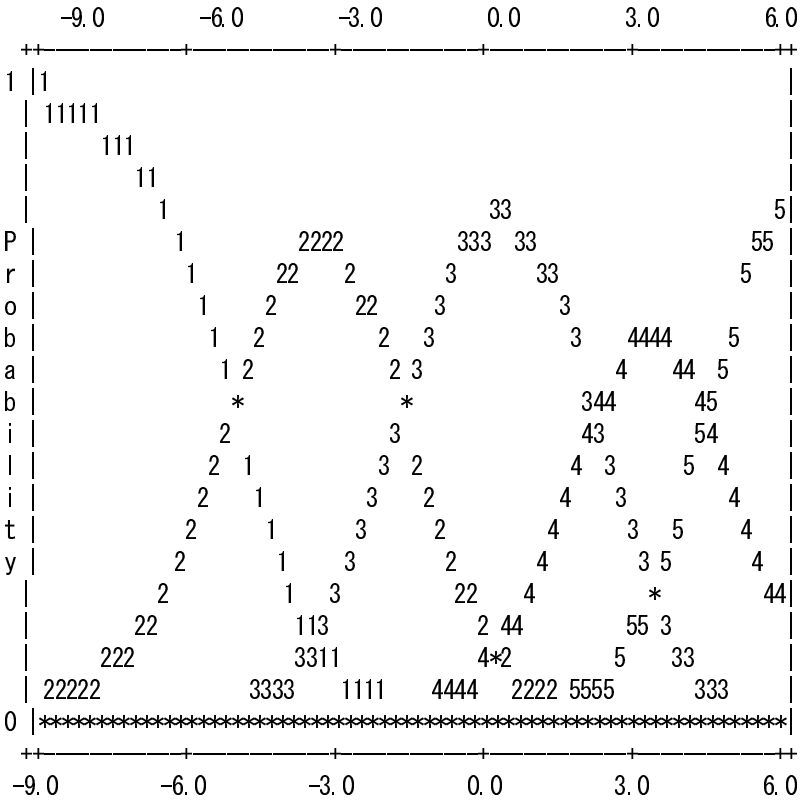


Table 11 shows the rating scale statistics for Impression. Average measures advance monotonically with each category. All outfit mean-squares are less than 2.0. All step difficulty increases fall within 1.4 and 5.0, which does meet (3). In Figure 6, the curves are like the expected succession of hills and obvious hill tops are observed, indicating that the step difficulties increase monotonically with rating scale numbers and the scales work as intended. In sum, the three rating scales conformed to expectations about its use.

5) Do individual raters score a particular group of subjects more harshly or more leniently? If so, what are the sub-patterns of ratings in terms of rater-subject interaction for each rater?

Table 12. Bias Calibration Report: Rater-subject Interaction for Rater 1

Subject	Ability (logits)	Observed score	Expected score	Obs-Exp Average	Bias (logits)	Error	z-score	Infit Mean Score
12	4.16	9	13.2	-1.41	-4.26	1.14	-3.72	0.0
19	0.76	12	9.7	0.77	2.25	0.96	2.35	0.0

Table 13. Bias calibration report: rater-subject interaction for Rater 3

Subject	Ability (logits)	Observed score	Expected score	Obs-Exp Average	Bias (logits)	Error	z-score	Infit Mean Score
4	-5.61	7	5.0	0.68	2.53	1.07	2.38	0.7

Table 14. Bias Calibration Report: Rater-subject Interaction for Rater 5

Subject	Ability (logits)	Observed score	Expected score	Obs-Exp Average	Bias (logits)	Error	z-score	Infit Mean Score
12	4.16	14	11.3	0.89	2.65	1.17	2.27	0.9
7	2.81	8	10.0	-0.65	-2.40	1.07	-2.24	0.9

Tables 12-14 show the results of the bias analysis in terms of interaction between rater severity and subject ability. Since the rater-subject interactions where z-score values fall below -2.0 or above 2.0 means a significant bias, only those interactions were listed in the tables. There were a total of five significantly biased interactions among Rater 1, Rater 3 and Rater 5. In each of Tables 12-14, the first two columns show subject ID (column 1) and the ability estimate for each subject (column 2). The next two columns show a total observed score (column 3) and a total expected score (column 4) of two tasks as well as the raters' impression on the subject. Since each scoring had a range of 1-5, the total observed or expected score falls in the range from 3 to 15. If a subject received a lower observed score from a rater than expected, the rater scored the subject more harshly than expected.

Column 5 shows the impressionistic scoring and the average difference between the total observed and expected scores from the rater for the subject across the two tasks. The next two columns show a bias logit, which presents the degree of the difference indicated in column 5 (column 6) and the likely error of the bias estimate (column 7). In column 8, the bias estimates in column 6 are converted into z-scores. A z-score below -2.0 indicates that the rater consistently scored the subject more leniently compared to how that particular rater scored the subject. Conversely, a z-score greater than +2.0 suggests that the rater consistently scored the subject more harshly than other subjects. In column 9, the infit mean square value shows how consistent the pattern of bias is for the rater to evaluate the subject's ability across the entire range of scoring. In this case, the mean of the infit mean square value was 0.4 and its standard deviation was 0.6. Thus, fit values above 1.6 logits suggest misfit (0.4+[0.6×2]).

In particular, raters' views on the rating of each task were considered to play important role in rater-subject interaction, so the three raters (Rater 1, 3 and 5) were asked to explain why they scored the subject more harshly/leniently than expected. Rater 1, who had scored Subjects 12 and 19 more harshly and leniently, respectively, as compared to how he scored the other subjects, explained his reasons as follows:

> As for Task 1 of Script 12, its organization as a letter of English is insufficient because there are neither salutations nor complimentary closings [Organizational skills]. The letter also exceeds the word limit [Number of words]. In Task 2, the content of item 2 is the same as item 3 as well as item 4 and 10 [Number of items]. The form used in items 5, 6 and 7 is all 'to get' [Communicative quality]. As for Task 1 of Script 19, the language is so concise and accurate that the readers easily understand the well-organized content [Linguistic accuracy]. In Task 2, the expression of infinitive phrases is rich in variation [Communicative quality] and its relevant ideas have a positive effect on the readers [Communicative effect].

Conversely, Rater 5, who had scored Subject 12 more leniently and Subject 7 more harshly, explained his reasons as follows:

> Task 1 of Script 12 demonstrates clear organization with a variety of linking devices [Organizational skills]. The writer explains one topic with some sentences so that the number words are sufficient [Number of words]. In Task 2, there are some similarities between items, but each of them is specific and relevant [Communicative effect]. On the contrary, task 1 of Script 7 lacks organization and development [Organizational skills]. The writer

> explains each topic with only one sentence. As for Task 2, there are many items written, but they are very similar in content [Number of items]. It shows limited choice of vocabulary to express the ideas [Communicative effect].

Rater 3, who had scored Subject 4 more leniently, explained his reasons as follows:

> The overall shape of Task 1 is hard to recognize [Organizational skills], but five of the twelve sentences focus on 'soccer,' so that message was clearly communicated to the reader [Communicative quality/effect].

The three raters' views are summarized in Tables 15 and 16, which indicate that each rater had a unique rater-subject bias pattern as follows:

> ・Rater 1 becomes harsher on accuracy when the number of words are exceeded and the script lacks an organizing principle and development. Rater 1 becomes more lenient when the script demonstrates linguistic accuracy. Rater 1 becomes harsher on communicability when the written items are similar, and its number is limited. Rater 1 becomes more lenient when the script displays adequate communicative effect.
> ・Rater 5 becomes more lenient on accuracy when there are a number of words and the script demonstrates clear organization with a variety of linking devices. Rater 5 becomes harsher when the script displays a lack of organizational skills. Rater 5 becomes harsher on communicability when the written items are similar, and its number is limited. Rater 5 becomes more lenient when the script displays adequate communicative effect.
> ・Rater 3 becomes harsher on accuracy when the script displays a lack of organizational skills, but becomes more lenient when the script demonstrates communicative effect.

Table 15. Raters' Views on the Rating of Task 1

Task 1	R1(S12)	R1(S19)	R3(S4)	R5(S12)	R5(S7)
Number of words	X			○	
Organizational skills	X		X	○	X
Linguistic accuracy		○			

Table 16. Raters' Views on the Rating of Task 2

Task 2	R1(S12)	R1(S19)	R3(S4)	R5(S12)	R5(S7)
Number of items	X			○	X
Communicative quality	X	○			
Communicative effect		○		○	X

6) Do the raters score particular tasks more harshly or more leniently than others? If so, what are the sub-patterns of ratings in terms of rater-task interaction for each rater?

Table 17 shows the results of the bias analysis in terms of the interaction between raters and tasks. It lists all rater-task interactions (5 raters×3 tasks) including ones without a significant bias. The first two columns show rater ID (column 1) and tasks (column 2). The next two columns show a total observed score of 20 subjects from the rater on each task (column 3) and a total expected score for 20 subjects from the rater on each task (column 4). Since the possible scores for each task fall in the range of 1-5 points, the total observed or expected scores of 20 subjects for each task falls in the range 20-100. If an observed score from a rater is higher than the expected score, the rater scored the subject more leniently than expected in the task.

Table 17. Bias Calibration Report: Rater-task Interaction

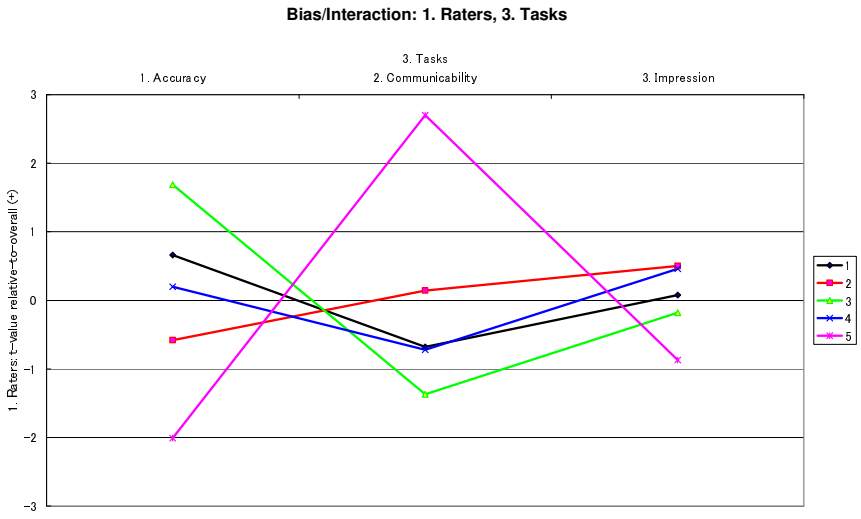
Rater	Tasks	Observed score	Expected score	Obs-Exp Average	Bias (logits)	Error	z-score	Infit Mean scorer
5	Communic ability	57	50.7	0.31	1.13	0.42	2.70	0.9
3	Accuracy	70	66.0	0.20	0.72	0.43	1.69	0.8
1	Accuracy	64	62.4	0.08	0.28	0.42	0.66	1.0
2	Impression	64	62.8	0.06	0.20	0.41	0.50	0.5
4	Impression	57	55.9	0.05	0.19	0.42	0.46	0.6
4	Accuracy	57	56.5	0.02	0.08	0.42	0.20	0.8
2	Communic ability	62	61.7	0.02	0.06	0.41	0.14	1.0
1	Impression	62	61.8	0.01	0.03	0.41	0.08	1.0
3	Impression	65	65.4	-0.02	-0.07	0.41	-0.18	0.6
2	Accuracy	62	63.4	-0.07	-0.24	0.42	-0.58	1.3
1	Communic ability	59	60.7	-0.08	-0.28	0.41	-0.68	0.9
4	Communic ability	53	54.7	-0.08	-0.31	0.43	-0.72	0.5
5	Impression	50	52.0	-0.10	-0.38	0.44	-0.87	0.7
3	Communic ability	61	64.3	-0.17	-0.56	0.41	-1.37	0.8
5	Accuracy	48	52.6	-0.23	-0.90	0.45	-2.01	1.3

Column 5 shows the average difference between the total observed and expected scores from the rater for the task across 20 subjects. The next two columns show a bias logit, which presents the degree of difference indicated in column 5 (column 6) and the likely error of the bias estimate (column 7). In column 8, the bias estimates in column 6 are converted into z-scores. A z-score below -2.0 indicates that the rater consistently scored the task more leniently compared to the way that particular rater scored other

tasks. Conversely, a z-score greater than +2.0 suggests that the rater consistently scored the task more harshly than others. In column 9, the infit mean square value shows how consistent the pattern of bias is for the rater to evaluate the task across all subjects. In this case, the mean of the infit mean square value was 0.8 and its standard deviation was 0.2. Thus, fit values above 1.2 logits suggest a misfit (0.8+[0.2×2]).

Table 17 shows that there were two interactions with a significant bias out of the entire 15 interactions. It also shows that the interactions that displayed a significant bias were distributed in one rater (Rater 5). Figure 7 plots graphically the information on rater-task interactions in the form of bias z-scores. This figure indicates that Rater 5 is harsher on communicability and more lenient on accuracy.

Figure 7. Rater-task bias/interaction



7) To what extent, statistically, is the task-based writing test a reliable and valid measure?

(1) Reliability

In the first analysis (Table 6) the data set was analyzed using FACETS. The table provided information on the characteristics of raters (severity and consistency). All raters displayed acceptable levels of consistency with themselves. This can be seen from the Infit Mean Square column, by adding two standard deviations to the mean. Raters falling within these parameters in their reported Infit Mean Square indices are considered to have behaved consistently. On the other hand, the separation and reliability figures indicate

that there were significant differences between raters in terms of severity. However, the difference, based on fair average scores, is 0.65 of one grade in the scale, suggesting that there would be no impact on scores awarded in an operational setting. The analysis of the two tasks and the impressionistic scoring in Table 8 show that no significant difference occurs between the tasks and the impressionistic scoring. The scoring forms do not appear to separate the subjects to a significant degree. This means that in normal operations the three scoring forms can be considered equivalent.

(2) Validity

In Table 8, an estimate of the item discrimination was computed according to a "Generalized Partial Credit Model" approach. 1.0 is the expected value, but discriminations in the range 0.5 to 1.5 provide a reasonable fit to the Rasch model (Linacre, 2007, p.132). All the estimates fall in this range (0.90, 1.05, 1.37), which indicates that the randomness in the three sets of data fit the Rasch model. The two tasks and the impressionistic scoring were, therefore, of relevance to dependent data acquisition.

Table 18 Inter-rater Correlation Coefficients between Raters' Scores and the *Criterion* Score

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Avg.
T1	.71	.70	.65	.63	.60	.66
T2	.74	.71	.67	.70	.79	.72
IS	.74	.78	.70	.72	.68	.72

Note. T1=task 1; T2=task 2; IS=impressionistic scoring

There is also evidence that detracts from the measure's validity. Table 18 shows the resulting correlation coefficients for the relationship between each of three raters' scores and the *Criterion* score, and they were statistically significant ($p<.01$) for Task 1, Task 2 and impressionistic scoring. This result supports the validity of the task-based writing test including these three scores.

5. Discussion

5.1 Summary

The results of the TBWT suggested that the students ability was effectively measured using these tasks and raters. The FACETS analysis showed that the difficulty of the two tasks and the impressionistic scoring were equivalent. The interrater correlation coefficients between pairs of raters were high, and the raters displayed acceptable levels of consistency with themselves. There were, however, significant differences between raters in terms of severity. A

bias analysis research was conducted in rater-subject interactions and in rater-task interactions. These analyses indicated that three of the five raters were significantly biased towards certain types of subjects, and these raters' bias patterns were unique. One of the three raters also had a clear pattern of significant interaction between the rater and specific tasks.

These findings suggest that the TBWT scoring guide may have contributed to the reduction of biased interactions, but training for certain raters with his/her unique bias patterns might still be required. It was assumed that the scoring guide gave raters a shared understanding of the construct of writing ability as defined by the test writers, and thus the scoring guide may effectively reduce the differences or biases caused by variation among raters. However, as previous research suggests, training and experience improve agreement among raters (Shohamy *et al*., 1992; Weigle, 1994). Lumley (2002) suggested that trained teacher raters garner the benefit of training by simply coping with the demanding task, shaping their natural impression to what they are required to do, and using the scale to frame the descriptions of their judgment of a text. This view of the function of training suggests that training plays an important role in influencing raters' behavior, so it may contribute to the variation in frequencies of biased interactions.

5.2 Implications

From the results of the present study using FACETS, three implications are drawn. First, five teacher raters were found to be self-consistent in scoring 20 different subjects' writing performance. However, there were relatively small but significant differences in overall rater severity. In addition, three of the five raters had a unique bias pattern toward a certain type of text. Fit statistics analysis of the raters in this study suggested that training for a certain rater with his/her unique bias pattern could have a major impact on rating behavior, meaning that the rater facet does not necessarily represent a problematic or validity-threatening part of the testing process.

Second, the 5-point scales were found to demonstrate acceptable fit, and seemed to be a more reliable tool in determining the estimate of subjects' writing ability. The scales associated with the five rating categories and their specific written samples were shown to be mostly comprehensible and usable by raters. However, it must be said that the raters in this study were all participants in the pre-testing. Raters tend to increase their internal consistency in assigning ratings as they gain experience (Weigle, 1998). Whether new teacher raters are self-consistent in scoring the same writing samples with the rating scales must be observed and confirmed in further studies.

Finally, one source of score variance in the writing performance test, task, was negligible in terms of difficulty. The assessment tasks used in this study provided reasonable fit to the Rasch model. This result implies that task

development based on the construct-based processing approach could be a reasonably solid basis to estimate students' writing ability, and those tasks may draw valid inferences to their writing performance.

6 Conclusion

In the present study, the results showed that the students' ability was effectively measured using the developed elicitation tasks and five teacher raters, and that all raters displayed acceptable levels of consistency with themselves. There were, however, relatively small but significant differences among raters in terms of severity. The bias analyses also indicated three of the five raters were significantly biased towards certain types of subjects, and these raters' bias patterns were unique. These findings suggest that the TBWT scoring guide may have contributed to the reduction of biased interactions, but training for certain raters with his/her unique bias patterns might still be required.

The FACETS analysis for this study showed that the difficulty of the two tasks and the impressionistic scoring were considered equivalent, which provided reasonable fit to the Rasch model. The equivalence of task difficulty may indicate that task development based on the construct-based processing approach could be reliable and valid to estimate students' writing ability. The rating scales associated with the five categories and their specific written samples were shown to be mostly comprehensible and usable by raters, and demonstrated acceptable fit. However, there is still room for argument about the reliability and validity of assessment tasks and rating scales.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests.* Oxford: Oxford University Press.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*, 453-476.

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice Hall Regents.

Burstein, J., Chodorow, M., & Leacock, C. (2003). Criterion^{SM} online essay evaluation: An application for automated evaluation of student essays. In J. Reid & R. Hill (Eds.), *Proceedings of the fifteenth annual conference on innovative applications of artificial intelligence* (pp. 3-10). Menlo Park, CA: AAAI Press.

Davies, A. (1990). *A principles of language testing*. Basil Blackwell.
Linacre, J. (2002). Guidelines for rating scales. *Mesa Research Note, 2*. Retrieved March 18, 2008, from http://www.rasch. org/rn2.htm
Linacre, J. (2007). *A user's guide to FACETS: Rasch-model computer program.* . Chicago, IL: MESA Press.
Linacre, J. (2008). *Facets*, version no. 3.63. Computer program. Chicago, IL: MESA Press.
Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*, 246-276.
Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.
Shohamy, E., Gordon, C., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal, 76*, 27-33.
Sugita, Y. (2008). *Task-based performance assessment of Japanese second language writing*, Paper presented at 42nd International Annual IATEFL Conference, Exeter, the UK, April.
Tyndall, B., & Kenyon, D. M. (1995). Validation of a new holistic rating scale using Rasch multifaceted analysis. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 39-57), Clevedon, England: Multilingual Matters.
Weigle, S. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*, 197-223.
Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*, 263-287.

Yoshihito Sugita
Yamanashi Prefectural University
1-6-1 Ikeda, Kofu City, Yamanashi,
400-0062, JAPAN
Tel & Fax: 81 55 253 8649
Email: sugita@yamanashi-ken.ac.jp

Received: January 7, 2009
Revised: May 19, 2009
Accepted: June 5, 2009

The Development and Implementation of Task-based Writing Performance Assessment for Japanese Learners of English

Appendix

A. Assessment tasks for pre-testing

1. Task 1 (accuracy)

- Rubric: This is a test of your ability to write a coherent and grammatically correct paragraph. You will have 20 minutes to complete the test.
- Prompt: You are going to stay with the Parker Family in Britain this summer. Write a 100-120 word letter introducing yourself to your host family. Before writing, think of the following topics:
  —Your name and age
  —Your job and major in school
  —Your family and pet
  —Your interests and hobbies
  —Your favorite places, foods and activities
  —Your experience traveling abroad
  —Some things you want to do while you are in Britain

2. Task 2 (communicability)

- Rubric: This is a test of your ability to write ideas relevant to the discussion topic without causing the reader any difficulties. You will have 10 minutes to complete the test.
- Prompt: You are going to discuss the following topic with your classmates, “Why do you study English?” In order to prepare for the discussion, think of as many answers as possible to the question and write them as “To travel abroad.”

Appendix B: Rating scales

[Accuracy]

Organizational skills	Linguistic accuracy
The writing displays a logical organizational structure which enables the content to be accurately acquired.	Errors of vocabulary, spelling, punctuation or grammar
The writing —is well organized and well developed (TWE) —shows strong rhetorical control	The writing —demonstrates appropriate word choice though it may have occasional errors (TWE)

and is well managed (M) —has clear organization with a variety of linking devices (FCE)	—has few errors of agreement, tense, number, word order/function, articles pronouns, prepositions, spelling, punctuation, capitalization, and paragraphing (ESL)
A (5) I strongly agree to assign the above criteria **B+**(4) I partially agree to assign the above criteria **B** (3) I agree to assign the above criteria **B**-(2) I disagree with assigning the above criteria **C** (1) I strongly disagree with assigning the above criteria	

[Communicability]

Communicative quality	Communicative effect
The writing displays an ability to communicate without causing the reader any difficulties The writing —displays consistent facility in use of the language (TWE) —contains well-chosen vocabulary to express the ideas and to carry out the intentions (M)	Quantity of ideas to develop a response and relevance of the content to the proposed task The writing —effectively addresses the writing task (TWE) —has a very positive effect on the target reader with adequately organized relevant ideas (FCE)
A (5) I strongly agree to assign the above criteria **B+**(4) I partially agree to assign the above criteria **B** (3) I agree to assign the above criteria **B**-(2) I disagree with assigning the above criteria **C** (1) I strongly disagree with assigning the above criteria	

Appendix C: Explanatory part of modified rating scales

「A(5) きわめてあてはまる」例

タスク１

Dear Parker Family,
Hello! My name is *** ***. Nice to meet you. I'm 19 years old
and a university student. There are 6 members in my family. They
are my father, mother, brother, sister, and grandmother and I. But
now I live alone in *** to study English education of junior high school
at *** university. I miss my family.
My hobbies are watching movies, listening to music, and playing

the clarinet. And I'm interested in world history.
I've been to America and Australia. Both of them were
homestay. They were great!
I'd like to talk much with you while I'm in England in order to
improve my English skills and to know your culture.
I'm looking forward to meeting you sooner.
Yours, ***** [125 words]

［解説］

・文章の構成および展開がうまくできている
・論理展開の方法が適切で説得力がある
・部分的に誤りはあるが、語彙使用が適切である
・主語と動詞の一致、時制、単数・複数、語順および語法、冠詞、代名詞、前置詞の使用にほとんど誤りがない
・スペル、句読法、大文字使用、段落分けの仕方にほとんど誤りがない

「A(5) きわめてあてはまる」例

タスク２

Discussion Topic: Why do you study English?

- To become an English teacher
- To talk with many people all over the world
- To be a good English speaker
- To make friends with foreigners
- To fall in love with foreigners
- To read Harry Potter
- To watch foreign movies that English is spoken
- To read English newspapers and magazines
- To send e-mail to my friend who lives in America
- To write a letter to my hostfamily who live in Australia
- To go the Desneyland which is in America
- To go shopping alone in New York

［解説］

・与えられた課題に対してそつ無く回答している
・読み手に対して非常に明瞭に内容が伝わる
・言語使用能力が確かなものであることがわかる
・自分の考えを表現したり、意図を伝えることのできるすぐれた語彙力がある