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

—PAAL 2009—

Yoshihito SUGITA

Introduction

- **The dual-mode system (Skehan, 2001)**
- **When time is pressing, and contextual support high, memory-based communication is appropriate.**
- **When there is more time, and precision is important, the rule-based system can be accessed.**

Developing a performance assessment

- **Construct-based approach (e.g., Alderson et al., 1995; Bachman & Palmer, 1996; Brown, 1996).**
- **Procedures for the 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.**

Construct Definitions (1): Accuracy

Accuracy (rule-based system, organizational knowledge)	
Organizational skills	Linguistic accuracy
The writing displays a logical organizational structure which enables the content to be accurately grasped.	Errors of vocabulary, spelling, punctuation or grammar.

Construct Definitions (2): Communicability

Communicability (exemplar based-system, pragmatic knowledge)	
Communicative quality	Communicative effect
The writing displays an ability to communicate without causing the reader any difficulties.	Quantity of ideas to develop the response and relevance of the content to the proposed task.

Purposes

- **In order to examine the degree of reliability and validity of the task-based writing performance test, the following are focused on:**
  - **(1) Raters' severity**
  - **(2) Interactions with writers' abilities and task difficulties**
  - **(3) Reliability of tasks and rating scales**
  - **(4) Measure's validity**

Test Participants and Materials

- **20 undergraduate students (6 males and 14 females), Native speakers of Japanese with an intermediate level of English proficiency**
- **Assessment tasks:**
  Task 1 focusing on accuracy (20 min.)
  Task 2 focusing on Communicability (10 min.)
- ***Criterion* Essay writing**

"Why do you think people attend college or university?" (30 min.)

Task 1

- **You are going to stay with 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, major in school**

Task 1

- **Your family and pets**
- **Your interests and hobbies**
- **Your favorite places, foods, activities**
- **Your experience in traveling abroad**
- **Some things you want to do while you are in Britain**

Task 2

- **You will have 10 minutes to make notes about the following discussion topic, "Why do you study English?" In order to prepare for the discussion, think of answers to the question as many as possible and write them as "To travel abroad."**

Scoring Materials and Procedure

- **Each of 40 scripts was scored by five experienced high school teachers**
- **Both scripts and scoring guidelines were given by mail**
- **The rating procedure:**

Task 1 → Task 2 → Total impression

Description of Accuracy (1)

Organizational skills
The written text -is well organized and well developed (TWE). -shows strong rhetorical control and is well managed (MWA). -has clear organization with a variety of linking devices (FCE).

Description of Accuracy (2)

Linguistic accuracy
The written text -demonstrates appropriate word choice though it may have occasional errors (TWE). -has few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions, spelling, punctuation, capitalization, paragraphing (ESL).

Description of Communicability (1)

Communicative quality
The written text -displays consistent facility in use of language (TWE). -contains well-chosen vocabulary to express the ideas and to carry out the intentions (MWA).

Description of Communicability (2)

Communicative effect
The written text -effectively addresses the writing task (TWE). -has a very positive effect on the target reader with adequately organized relevant ideas (FCE).

5-point Likert Scale for Rating

A (5) : I strongly agree to assign the above description

B+ (4) : I partially agree to assign the above description

B (3) : I agree to assign the above description

B- (2) : I disagree with assigning the above description

C (1) : I strongly disagree to assign the above description

Data Analysis

- **The data were analyzed using FACETS (Linacre, 2008).**
- **To examine the measurement characteristics of the testing, three facets were specified: subject, rater and task.**
- **Bias analysis: Rater × Subjects**
  Rater × Tasks

Result (1) : Raters

Raters	1	2	3	4	5
Severity	0.35	0.52	0.97	-0.69	-1.41
Error	0.24	0.24	0.24	0.24	0.25
Infit	0.97	0.93	0.78	0.65	1.16

→There was a significant difference in severity among raters, but all raters behaved consistently in the scoring.

Result (1) : Bias analysis of Rater 1

Subject	Observed score	Expected score	Obs-Exp	Bias(logits)	Z-score
12 (U)	9	13.2	-1.41	-4.26	-3.72
19 (M)	12	9.7	0.77	2.25	2.35

Accuracy: excessive words, lack of an organizing principle and development in script → harsher
linguistic accuracy → more lenient
Communicability: similar items, limited number of items → harsher
Adequate communicative effect → more lenient

Result (1) : Bias analysis of Rater 5

Subject	Observed score	Expected score	Obs-Exp	Bias(logits)	Z-score
12 (U)	14	11.3	0.89	2.65	2.27
7 (M)	8	10.0	-0.65	-2.40	-2.24

Accuracy: lack of organizational skills → harsher
many words, linking devices → more lenient
Communicability: similar items, limited number of items → harsher
Adequate communicative effect → more lenient

Result (1) : Bias analysis of Rater 3

Subject	Observed score	Expected score	Obs-Exp	Bias(logits)	Z-score
4 (L)	7	5.0	0.68	2.53	2.38

Accuracy: lack of organizational skills → harsher t
Communicability: adequate communicative effect → more lenient

Result (2) : Task Difficulty

	Task 1	Task 2	Impression
Difficulty	0.13	-0.18	0.50
Error	0.19	0.19	0.19
Infit	1.10	0.92	0.68
Discrimination	0.90	1.05	1.37

→ No significant variation in difficulty exists among the tasks and impressionistic scoring, $\chi^2(2)=1.5$, p=.47
→Estimate of Discrimination: 0.5 < E.D. < 1.5, reasonable fit with the Rasch model

Result (2): Reliability (accuracy)

Rater	Observed score	Expected score	Obs-Exp	Bias (logits)	Z-score
1	64	62.4	.08	.28	.66
2	62	63.4	-.07	-.24	-.58
3	70	66.0	.20	.72	1.69
4	57	56.5	.02	.08	.20
5	48	52.6	-.23	-.90	-2.01

→ Rater 5 consistently scored the task more leniently, but Raters 1-4 evaluated the task without the pattern of bias across all subjects.

Result (2): Reliability (communicability)

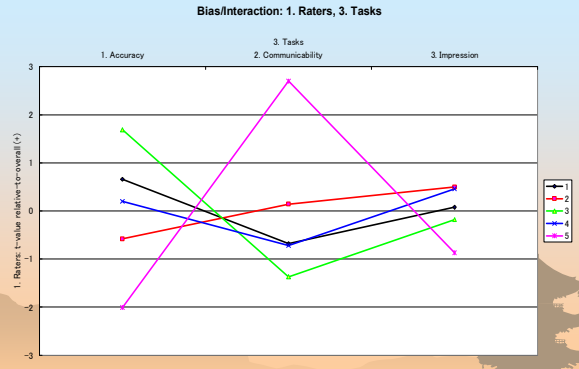
Rater	Observed score	Expected score	Obs-Exp	Bias (logits)	Z-score
1	59	60.7	-.08	.28	-.68
2	62	61.7	.02	.06	.14
3	61	64.3	-.17	-.56	-1.37
4	53	54.7	-.08	-.31	-.72
5	57	50.7	.31	1.13	2.70

→ Rater 5 consistently scored the task more harshly, but Raters 1-4 evaluated the task without the pattern of bias across all subjects.

Result (2): Reliability (Impression)

Rater	Observed score	Expected score	Obs-Exp	Bias (logits)	Z-score
1	62	61.8	.01	.03	.08
2	64	62.8	.06	.20	.50
3	65	65.4	-.02	-.07	-.18
4	57	55.9	.05	.19	.46
5	50	52.0	-.10	-.38	-.87

→ All raters scored holistically without the pattern of bias across all subjects.



Result (2) : Validity

	R1	R2	R3	R4	R5	Av.
Task 1	.71	.70	.65	.63	.60	.66
Task 2	.74	.71	.67	.70	.79	.72
Impression	.74	.78	.70	.72	.68	.72

Each of three raters' scores and the Criterion score were statistically significant (p<.01)

Result (3): Rating scale (Accuracy)

Category	N	%	STEP	Outfit
1	10	10		.9
2	23	23	-5.28	1.1
3	33	33	-1.54	1.4
4	24	24	1.96	.8
5	10	10	4.86	1.0

All outfit mean-squares are less than 2.0, all increases in step difficulty fall within 1.4 and 5.0.

Result (3): Rating scale (Communicability)

Category	N	%	STEP	Outfit
1	13	13		1.2
2	23	23	-4.97	1.6
3	33	33	-1.49	.9
4	21	21	1.98	.6
5	10	10	4.49	.7

All outfit mean-squares are less than 2.0, all increases in step difficulty fall within 1.4 and 5.0.

Result (3): Rating scale (Impression)

Category	N	%	STEP	Outfit
1	11	11		.7
2	22	22	-5.13	.5
3	36	36	-1.65	.7
4	20	20	2.27	.7
5	11	11	4.51	.8

All outfit mean-squares are less than 2.0, all increases in step difficulty fall within 1.4 and 5.0.

Conclusion

- All raters displayed acceptable levels of consistency, but there were relatively small but significant differences among raters.
- The difficulty of the two tasks and impressionistic scoring were considered equivalent.
- The rating scales mostly comprehensible and usable by raters, and demonstrated acceptable fit.

Implications

- Three of the five raters were significantly biased towards certain types of subjects.
- The raters' bias patterns were unique.
- The question of 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.