Comparability of Accuracy and Communicability Tasks: Are they all Equally Difficult?

Yoshihito Sugita

Faculty of Nursing, Yamanashi Prefectural University

sugita@yamanashi-ken.ac.jp

Abstract

In this study, two sets of accuracy and communicability tasks (*Original* and *New*) are compared to determine whether these tasks are parallel in terms of task difficulty. Three experienced Japanese teachers of English examined the degree of parallelness of five accuracy and five communicability tasks including *Original* tasks with regard to time pressure, support, stakes and overall difficulty, and selected comparable tasks (*New* tasks). *Original* and *New* tasks were administered to 24 university students in Japan. The two raters were requested to mark each of the two task sets seperately after an interval of one month. The performance data from the different task sets were analyzed using classical procedures (correlation and ANOVA) and FACETS. The results showed that *Original* and *New* tasks can be considered parallel at the overall test level. At the individual task level, the two communicability tasks could be thought of equally difficult, while the accuracy tasks are fairly reliably separated into different levels of difficulty. Reasons why the degree of difficulty of the selected accuracy task varies were suggested by prompt effects in writing performance assessment, and provided useful insights to the further task-development.

Keywords

Writing, assessment, task difficulty

Introduction

With the aim of offering Japanese learners of English a reliable writing performance test, assessment tasks for accuracy and communicability were newly designed and developed. In order to provide test-takers with a fair writing test, it was thought necessary to establish parallelness of test tasks. The focus for this present study is whether two sets of accuracy and communicability tasks (*Original* and *New*) are parallel in terms of task difficulty.

1 The Study

1.1 Objectives

The objective is to investigate the parallelness of comparable tasks and to explore how equally difficult assessment tasks can be constructed.

1.2 Participants

1.2.1 Task examiners

The task examiners were three experienced Japanese teachers of English. They learned how to rate a task-based writing performance test (TBWT) using accuracy and communicability rating scales, and had an experience of rating *Original* tasks in 2010.

1.2.2 Test-takers

The test-takers were 24 second-year students from the faculty of nursing. All of the test-takers were native speakers of Japanese with an intermediate level of English language proficiency.

1.2.3 Raters

The two raters were female and native speakers of Japanese. Both of them were at the time of study teaching General English to undergraduates of diverse disciplines at various universities. They received training session before rating, in which they had prior experience with the rating scales and the assessment tasks. They shared similar backgrounds in terms of qualifications of more than twenty years of teaching experience.

1.3 Data collection and analysis

1.3.1 Data collection

The three task examiners investigated the degree of parallelness of five accuracy and five communicability tasks including *Original* tasks with regard to time pressure, support, stakes and overall difficulty, and selected comparable tasks (*New* tasks). *Original* and *New* tasks were administered to 24 university students in Japan. The two raters were requested to mark each of the two task sets seperately after an interval of one month.

1.3.2 Data analysis

The performance data from the different task sets were analyzed using classical procedures (correlation and ANOVA) and FACETS.

2 Results

2.1 Descriptive statistics

Tables 1 and 2 report for each scoring in *Original* and *New* tasks, its mean and standard deviation by the two raters (Rater 1 and Rater 2). It can be seen that in each task type the mean scores for all ratings are relatively close, ranging from 3.00 to 3.38 (Accuracy 1 and 2), 3.33 to 3.67 (Communicability 1 and 2) and 3.08 to 3.38 (Impression 1 and 2).

Table 1 Descriptive statistics of scoring *Original* tasks

Tasks 1	Accuracy		Communicability		Impression	
Raters	R1	R2	R1	R2	R1	R2
Mean	3.38	3.21	3.42	3.33	3.33	3.21
SD	0.75	0.86	0.95	0.79	0.80	0.82
Max.	5	5	5	5	5	5
Min.	2	2	2	2	2	2

Table 2 Descriptive statistics of scoring *New* tasks

Tasks 2	Accuracy		Communicability		Impression	
Raters	R1	R2	R1	R2	R1	R2
Mean	3.21	3.00	3.67	3.33	3.38	3.08
SD	1.04	1.15	0.75	0.80	1.07	0.86
Max.	5	5	5	5	5	5
Min.	2	1	3	2	2	2

2.2 Correlation

For equivalent tasks in *Original* and *New* tasks, the Spearman correlation coefficients fall in a range of .562 to .913, which are all significant at the 0.01 level. At the task set level (i.e. adding together scores on each task within sets), the correlation coefficient was .83, meaning strong relationship between the scores of the two task sets.

2.3 ANOVA

There was no significant differences between *Original* and *New* tasks, but the differences among the two tasks and impressionistic scores were recognized. A *post hoc* Bonferroni test indicated that there was no significant differences between any of the pairs of equivalent tasks. There were, however, significant differences between the accuracy and communicability tasks (*Original*: $p = .0005$; *New*: $p = .0016$), and between the communicability task and impressionistic scoring (*Original*: $p = .0134$; *New*: $p = .0134$).

2.4 FACETS analysis

FACETS analyses on the actual two tasks (Accuracy and Communicability) were conducted to investigate whether they might be considered equivalent. *Original* and *New* tasks do not appear to separate the subjects to a significant degree. According to the fixed chi-square test, there was no significant difference between each composite score across the two tasks in the two task sets ($p = .52$), indicating the two task sets can be considered equivalent.

At the individual task level, the results show that the mean of *Original a*nd *New* tasks calibrated on the logit scale were very close, with a difference of just .18. For *New* tasks, there is only a difference of .43 between the logit values for Accuracy 2 and Communicability 2, showing that these two tasks were similar in difficulty. To confirm this finding, the fixed chi-square test failed to reject the null hypothesis that the two types of *New* tasks could be thought of equally difficult ($p = .32$) .

Between the corresponding tasks in the two sets, the logit value for Communicability 1 and 2 is a difference of .70, as compared to a difference of 1.08 between the logit value for Accuracy 1 and 2. FACETS analyses on the pairs of equivalent tasks were also repeated to investigated whether they could be thought of equally difficult. In this analysis, the reliability is .69 for the accuracy tasks and .19 for the communicability tasks, indicating that the accuracy tasks are fairly reliably separated into different levels of difficulty, but the communicability tasks are not. The fixed chi-square test failed to reject the null hypothesis ($p = .12$) ; in other words, the two communicability tasks could be thought of equally difficult.

3 Discussion and conclusion

In sum, *Original a*nd *New* tasks can be considered parallel at the overall test level. At the individual task level, the two communicability tasks could be thought of equally difficult, while the accuracy tasks are fairly reliably separated into different levels of difficulty. Reasons why the degree of difficulty of the selected accuracy task varies were suggested by prompt effects in writing performance assessment (Slater & Mickan, 2001; Weigle, 2002), and provided useful insights to the further task-development.

References

Slater, S., & Mickan, P. (2001). Response validity and writing: a qualitative investigation of an international test of English. *Sheian Jogakuin University Journal*. 1, 165-183.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: CUP.