二つのパフォーマンステストの果たす役割
Complementary Role of Two Performance Tests in Schools

中村　優治 NAKAMURA, Yuji
● 東京経済大学
Tokyo Keizai University

Keywords パフォーマンス評価，ライテイングテスト，スピーキングテスト，評価者
Performance assessment, writing test , speaking test, raters

ABSTRACT

パフォーマンステストは受験者にスピーキングテストやライティングテストを実施して言語表出能力を測定するものである．いずれのテストにおいても受験者が具体的な作業を行ってその運用能力を測定しているため，テスト実施者はそれぞれのテスト結果をパフォーマンステストとしてひとつの枠組みの能力として捉える傾向にある．本稿の目的は，果たして受験者の言語表出能力は2種類のテストにおいてどのような結果を示すのか実験によって検証を試みようとするものである．具体的には日本人学生の言語表出能力（スピーキング能力とライティング能力）の類似点・相違点をパフォーマンステスト結果の分析を通して行い，分析には項目応答分析のファセットプログラムを用い，評価者間の信頼性の検証も行っている．結果の一部から，学生のパフォーマンス能力のより妥当性があり，信頼性の高い測定のためにはこの2種類のテストはいずれも欠くことができないと思われる．

1 Theoretical background and rationale

When we talk about performance assessment, we must first be clear about what performance is. Performance is the behavior exhibited by a test candidate in completing a particular task, a ratable sample of language. (Davies et al 1999) While the assessment of ability is based on this observable behavior, it is recognized that aspects of the testing situation may cause the candidate to perform in a way that does not allow an accurate measure of his/her ability to be obtained. (Davies et al 1999)

Gipps (1994) says that performance assessment is a term currently in wide use by people who want to move away from traditional standardized multiple-choice testing. She also states that performance assessment intends to model the real learning activities that teachers want students to engage with, oral and written communication skills, problems solving activities and so on.

Hambleton (1996) claims that performance tests should use direct methods of assessment (e.g. writing samples to assess writing, and oral presentations to assess speaking skills). He further says that performance tests should have a high degree of realism about then (that is "fidelity" should be high).

Milanovic (1998) defines a performance test as a test procedure which requires the candidate to produce a sample of language, either in writing or speech (e.g. essays and oral interview).

Performance tests that require test takers to produce a written or spoken sample of language can be collectively categorized as a single type of performance assessment.

However, results produced by test takers in these two different types of tests may vary. In recent years, researchers in second language acquisition and language testing have investigated in the influence of a range of task conditions and task characteristics on language test performance (Brindley 2002). He also states that many studies have revealed considerable variability in learners' spoken or written production according to the type of elicitation tasks (Brindley 2002).

2 Purpose of the research

The present research attempts to explore the similarities and differences among students as well as among evaluation items in two kinds of performance tests (Writing and Speaking) by using the Many-Faceted Rasch model, which is reputedly a powerful tool for handling polychotomous data involving raters' judgments.

3 Research design and methods

The subjects were 32 Japanese college students majoring in Law. Their tasks were 1) to write a speech presentation manuscript as a Writing Test and 2) to give, in class, an oral presentation based on the manuscript as a Speaking Test, which was tape-recorded for later evaluation. Four raters judged the 32 students' written manuscripts and oral presentation tapes using a 4-point scale. For the rating, 7 items were used for Writing (Grammar, Vocabulary, Fluency, Content, Discourse, Organization and Overall) and 8 items for Speaking (the same seven items as for Writing, plus pronunciation).

3. 1 Writing Test design

Subjects: 32 university students

Task: Each student wrote a composition on a single topic that he/she chose.

Raters: 4 raters (Three native speakers of English – A,B,C, and one non-native speaker of English--D)

Paired raters: 6 pairs (AB, AC, AD, BC,BD, CD)

Rating: Each student was rated by two raters.

Items: 7 evaluation items (Grammar, Vocabulary,

Discourse, Fluency, Content, Organization, Overall)

N.B. Discourse=Logicality, Content = Originality, Fluency = A measure of how easy the paper was to read, Overall = General Impression

Rating scale: 4-point scale (1=poor, 2, 3, 4=good)

3. 2 Speaking Test design

Subjects: 32 university students

Task: Each student gave an oral presentation in class based on the manuscript he/she wrote.

Raters: 4 raters (the same raters as above in the Writing Test)

Rating: Each student was rated by 4 raters.

Items: 8 evaluation items (Grammar, Vocabulary, Discourse, Fluency, Content, Organization, Overall, Pronunciation)

N.B. Discourse = Logicality, Content = Originality, Fluency=A measure of how easy it was for the student to speak, Overall=General Impression

Rating scale: 4-point scale (1=poor, 2, 3, 4 =good)

4 Data analysis

The data were analyzed using the Many-Faceted Rasch Measurement Model, which was able to give detailed information about three facets (student ability, item difficulty, and rater severity). The benchmark for the acceptable range of the infit and outfit statistics was set between 0.6-1.4 since this was performance test polytomous data that involved raters' judgments.

5 Results and discussion

Before comparing rating items and students in the two tests, we will first examine the raters' consistency in each test by looking at the fit statistics in Table 1 and Table 2. This investigation of the raters' consistency is important since it is the basis of comparing the other two facets of the test results (items and students).

On the basis of this rater investigation, we will discuss the results of the comparison of rating items in the two tests as well as the results of the comparison of students in the two tests.

5. 1. Investigation of raters' fit statistics in the two tests (the Writing Test and the Speaking Test)

Table 1 demonstrates that there are no misfitting raters, as is shown in MNSQs of the Infit-Outfit statistics columns. All the raters are within the acceptable range (0.6-1.4), which is usually applicable to a writing and speaking performance test rating scale. They function quite well within this group. In other words, we can count on their inter-rater reliability.

Table 1 Raters Measurement Report (in Writing Test)

Obsvd	Obsvd	Obsvd	Fair-M	Model		Infit		Outfit		
Score	Count	Average	Avrage	Measure	S.E.	MnSq	ZStd	MnSq	ZStd	N raters
309	112	2.8	2.85	-.76	.17	.9	0	1.0	0	1 A
314	112	2.8	2.79	-.99	.17	.9	0	.9	0	2 B
365	126	2.9	2.98	-.22	.16	1.1	0	1.1	0	3 C
348	98	3.6	3.54	1.97	.23	.7	-1	.9	0	4 D

Table 2, in the column of the Infit-Outfit statistics, demonstrates that there are no misfitting raters, i.e. all the MNSQs are within the range of 0.6-1.4. As a whole, each individual rater works well within the acceptable range of this group of four. This table also indicates that the exact agreement of raters was 49.6%, which is rather high as a performance test rater judgement.

Since Tables 1 and 2 provide us with convincing evidence of the raters' reliable performance in the two tests, we will now be able to go further into the comparison of the results of the rating of the items and of the students in the two tests.

5. 2 Comparison of the rating items

Table 3, along with the graphical description of Figure 1 below, gives the following results:

The correlation coefficient between the two groups (Writing Test rating items and Speaking Test rating items) is . 71 (and . 93 without grammar, see N.B. below), which is a rather high positive correlation for a performance test. In other words, except for pronunciation, there is a 50% overlap between the items on the two tests (the variance of .71=.49, i.e., a 50% overlap between the two tests).

As a whole, content and vocabulary are rather easy for the students, followed by organization and discourse, while fluency is rather difficult. Grammar is by far the most difficult rating item in Writing, whereas in Speaking it is in the mid-difficulty range. Pronunciation, which is only applicable to the Speaking Test, is also a difficult item for students.

Putting it another way, raters are more lenient about content or vocabulary in both tests, while they are more severe about grammar in Writing and about pronunciation in Speaking.

Although it is difficult to make a generalization from this small sample, it is still evident that a pronunciation test should be given in order to measure speaking ability more precisely, and an additional grammar test could be given to assess writing ability more accurately. In other items, such as content or vocabulary, we can make an overall prediction of item functioning from the results of just one type of test (i.e., from a Speaking Test to a Writing Test or vice versa).

In summary, these data show that in most cases the raters used the items in a similar way in order to evaluate students' ability in the two tests and that the items function more or less in a similar way in the two tests, though grammar and pronunciation are rather exceptional.

Table 2 Rater Measurement Report (in Speaking Test)

Obsvd	Obsvd	Obsvd	Fair-M	Model		Infit		Outfit		
Score	Count	Average	Avrage	Measure	S.E.	MnSq	ZStd	MnSq	ZStd	N raters
703	256	2.7	2.78	-.39	.12	1.2	2	1.2	2	1 Anthony
810	256	3.2	3.16	1.24	.13	1.1	0	1.1	0	4 Yuji
698	256	2.7	2.76	-.46	.12	1.0	0	1.0	0	2 David
702	256	2.7	2.77	-.40	.1	.7	-3	.7	-3	3 Mimy

Rater Agreement opportunity: 1536 Exact agreements: 762=49.6% Expected:777.1=50.6%

Table 3 Comparison of rating items

	writing	speaking
grammar	-1.02	0.03
discourse	0.22	0.44
content	0.64	0.98
vocabulary	0.58	0.53
fluency	-0.34	-0.61
organization	0.16	0.5
overall	-0.23	-0.19
pronunciation	0	-1.68

N.B. The Correlation coefficient=.71 (.93 without grammar, which is statistically quite far from the other items, especially in the Writing Test). Pronunciation should naturally be omitted in the correlation statistics because there is no equivalence in the Writing Test.

5. 3. Comparison of students

Table 4, along with the graphical description of Figure 2, gives the following results:

The correlation coefficient of the students' ability between the two tests (Writing and Speaking) is .50 (the variance is .25), which is quite low, although this is still a positive coefficient. There is only a 25% overlap between the two test results concerning student ability. In other words, on the whole, it is rather difficult to predict, with certainty, students' writing ability from speaking ability or vice versa, even if both of the tests are categorized in a performance test.

Figure 2 provides us with roughly 3 types of proficiency student groups: 1) those who are better at speaking, 2) those who show a positive correlation between speaking and writing, and 3) those who are better at writing. This result can be used for placement purposes where Group A needs improvement in writing skills, Group B can be taught both writing and speaking in equal balance, and Group C needs improvement in speaking skills. In other words, these data indicate that two different performance tests are necessary for reciprocal purposes to measure different language level students.

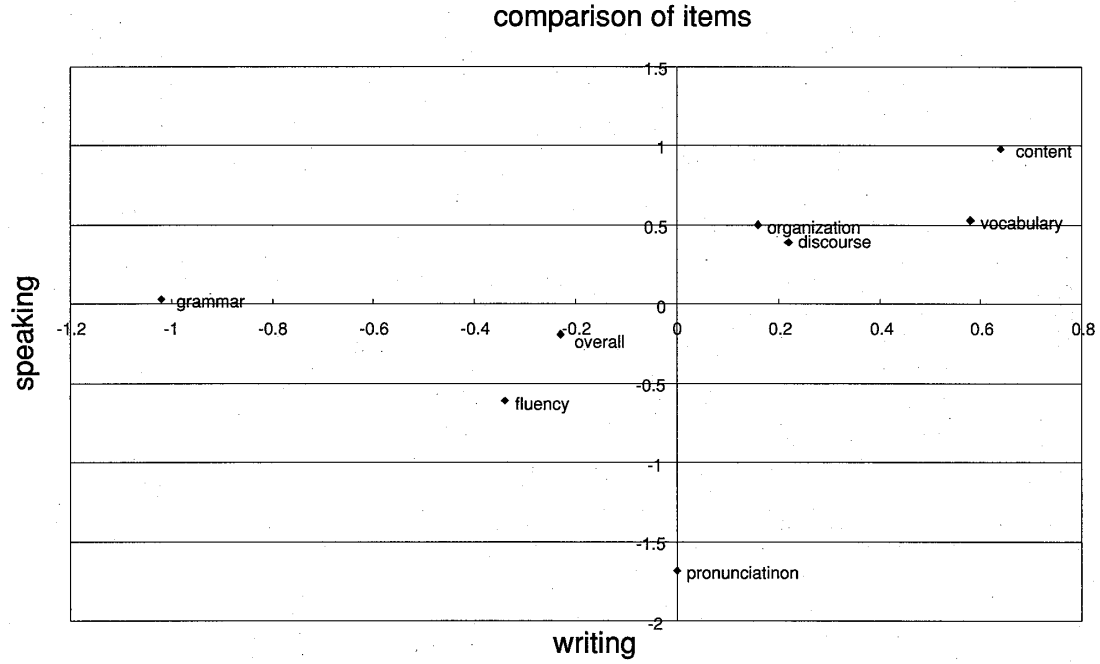

Figure 1

Table 4 Comparison of students

	writing	speaking
1	-0.44	0.29
2	1.21	-0.13
3	1.46	-0.45
4	1.46	-0.34
5	4.84	0.13
6	2.48	0.08
7	0.85	1.19
8	-0.02	0.62
9	-0.22	0.73
10	-0.22	0.08
11	-0.42	-0.03
12	2.25	1.55
13	1.95	2.03
14	3.05	1.31
15	1.13	0.18
16	3.28	1.67
17	2.59	1.08
18	3.98	1.31
19	2.02	1.79
20	1.21	2.53
21	0.51	0.51
22	0.97	1.55
23	3.83	2.77
24	0.06	-0.13
25	2.36	2.16
26	-0.17	0.51
27	3.04	3.02
28	-1.06	-0.24
29	3.7	1.55
30	-1.06	0.73
31	2.6	3.94
32	3.7	4.39

N.B. correlation coefficient= .50.

6 Conclusions

In terms of items, grammar is additionally necessary for measuring students' writing ability more precisely, and a pronunciation assessment is also necessary for a more accurate evaluation of speaking ability. In other items, information from the two tests (the Speaking Test and the Writing Test) can be shared rather reliably.

In terms of students, this study showed that students could be divided into three categories: 1) those who were better at speaking, 2) those who were better at writing and 3) those who showed a positive correlation between writing and speaking ability. The study also showed that in grammar students were weaker in writing than in speaking, which leads us to the conclusion that we need two performance tests (Speaking and Writing).

Finally, the present research on the two types of performance tests demonstrates the need for two-dimensional performance assessment involving both speaking and writing. This type of assessment makes it possible for student performance and ability to be measured more accurately and enables

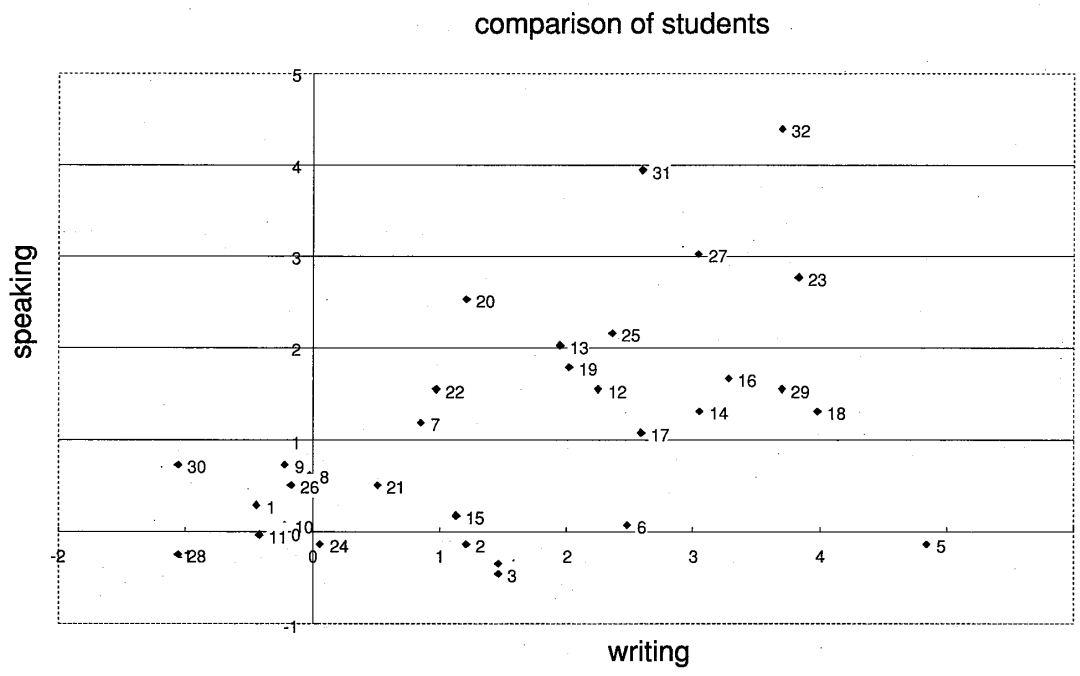

Figure 2

raters and teachers to make judgments that more precisely reflect students' performance in tests.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika* 43: 561-573.

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

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

Bode, R. K., & Wright, B. D. (1999). Rasch measurement in higher education. *Reprinted from Higher Education: Handbook of Theory and Research*, Vol. XIV (pp. 287-316). New York: Agathon Press.

Bonk, W.J. & Ockey, G.J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20,1*,89-110.

Brindley, G. (2002). Issues in language performance assessment. In Kaplan R.B.(Ed.) *The Oxford handbook of applied linguistics*. Oxford: Oxford University Press.

Brown, J. D. (1996). *Testing in language programs*. N.J.: Prentice Hall.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 201*, 1-25.

Davies, A., Brown, A. , et al. (Eds.). (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

Davies, A. & Elder, C. (Eds.). (2004). *The handbook of applied linguistics*. Oxford: Blackwell.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge, U.K.: Cambridge University Press.

Fulcher, G. & Reiter. R.M.,(2003). Task difficulty in speaking tests. *Language Testing, 20, 3*, 321-344.

Gipps, C.V. (1994). *Beyond testing*. London: The Falmer Press.

Hambleton, R.H. (1996). Advances in assessment models, methods, and practices. In Berliner, D.C. & Calfee R.C.(Eds.). *Handbook of educational psychology* (pp. 899-925). New York: Macmillan.

Hughes, A. (2003). *Testing for language teachers*. Second Edition. Cambridge: Cambridge University Press.

Kaplan, R. (Ed.). (2002). *The Oxford handbook of applied linguistics*. Oxford: Oxford University Press.

Linacre, J. M. (1989, 1993, 1994). *Many-facet: rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (1998). A *User's guide to facets rasch measurement computer program*. Chicago, IL: MESA Press.

Linacre, J. M. and Wright, B. D. (1998). *Facets: many-faceted rasch analysis*. Chicago, IL: MESA Press.

Linacre, John M. & Wright, B. (1998). *A User's Guide to Bigsteps/Winsteps: Rasch-Model Computer Program*. Chicago, IL: MESA Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika 47*: 149-174.

McNamara, T. (1996). *Measuring Second Language Performance*. London : Longman.

McNamara, T. (2004). Language Testing. In A. Davies & C. Elder (Eds.), *The Handbook of applied linguistics* (pp.763-781) Oxford: Blackwell

Milanovic, M. (1998). *Studies in language testing 6: Multilingual glossary of language testing terms*. Cambridge, UK: Cambridge University Press.

O'Loughlin, K. (2001). The equivalence of direct and semi-direct speaking tests: *Studies in language testing 13*. Cambridge, U.K.: Cambridge University Press.

Rasch, G. (1960, 1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen and Chicago: University of Chicago Press.

Tatum, D. S. (1997). *The results of a meaningful measurement analysis*. Chicago: National Communication Association.

Tatum, D. S. (1999). *Persuasive communication*. Chicago: The University of Chicago, Center for Continuing Studies

Wright, B. D. (1997). A history of social science. *Educational measurement: issues and practice. (Winter 1997)*, 33-45 & 52.

Wright, B. D. (1997). Fundamental measurement for psychology. In Emretson, S. and Hershberger S. (Eds.). *The new roles of measurement: What every psychologist and educator should know* (pp. 65-104). Hillsdale NJ: Lawrence Erlbaum Associates.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: rasch measurement*. Chicago, IL: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design: rasch measurement*. Chicago, IL: MESA Press.

Note

1 This paper is based on the presentation at the LTRC 2004 Convention in Temecula, California, March 26-28, 2004.

2 This research was supported in part by Tokyo Keizai University under Research Grant (1B03-04 and 2B04-05).