対話テストと討論テスト
Dialogue Interview Test and Multilogue DiscussionTest

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

Keywords スピーキングテスト，対話テスト，討論テスト，評価スケール
Speaking Test, Dialogue Test, Multilogue Test, Rating Categories

ABSTRACT

本研究は日本人大学生のスピーキング能力を対話テスト及び討論テストによって測定しようと試みたものである。学習者の社会言語能力が対話あるいは討論でどのように測定され，またどのような評価項目が有効なのかということを主に考察する．

1. Theoretical background and rationale

One of the major objectives of teaching oral communication is enhancing students'ability to use oral language in various sociolinguistic contexts. Speaking is often interactive, involving more than one person at the same time. Of course, speaking can be monologic, involving only one speaker, such as a lecture or a radio broadcast. Even if we limit the contexts to academic settings, there are various situations where students perform differently. Some students are good at monologue type speech making tests, others are skillful in handling dialogue type interview tests. Still some others are active in discussion activities (cf. Brown 2003; Bonk 2003).

2. Purpose of the research

This paper examines students' oral performance in two types of speaking tests, a dialogue interview test and a multilogue discussion test where conversational interactions between the speakers are inevitable and sociolinguistically appropriate language use is required. This paper also investigates the effectiveness of the ten evaluation items (1: Dialogue Grammar, 2: Dialogue Fluency, 3: Dialogue Vocabulary, 4: Dialogue Conversation Strategies, 5: Dialogue Sound System, 6: Multilogue Grammar, 7: Multilogue Fluency, 8: Multilogue Vocabulary, 9: Multilogue Conversation Stra-tegies, 10: Multilogue Content) and the rater's characteristic.

Research Question:
What do the Dialogue and Multilogue Tests results tell us about the following ?

a. the relationship between the two facets (students, items)
b. student ability
c. item difficulty
d. the construction of items in order of difficulty
e. the function of rating categories
f. possible construct of speaking
g. comparison of students' test results between Dialogue test and Multilogue Test

3. Research design and method

46 students took both the dialogue (interview) test and the multilogue (discussion) test. In the interview test students were interviewed by the classroom teacher and in the discussion test students were divided into groups of three or four members and they themselves conducted the discussion. Both data were evaluated by a classroom teacher using a four-point scale in five categories in each test. The data were analysed using the Rasch model.

3. 1 Dialogue Test design

Subjects: 46 university students
Task: Each student took an interview testconducted by the classroom teacher in the classroom setting.
Rater: Classroom teacher
Items: 5 evaluation items
Dialogue Grammar,
Dialogue Fluency,
Dialogue Vocabulary,
Dialogue Conversation Strategies,
Dialogue Sound System (Pronunciation)
Rating scale: 4-point scale (1, 2, 3, 4)

3. 2 Multilogue Test design

Subjects: 46 university students (the same students as above)
Task: Students made groups (consistig

of 3-4 people each) and discussed some given topics.

Raters: Classroom teacher

Items: 5 evaluation items

Multilogue Grammar,
Multilogue Fluency,
Multilogue Vocabulary,
Multilogue Conversation Strategies,
Multilogue Content)

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

4. Results and Discussion (See tables and graphs)

Table 1 suggests that in the column of the infit-outfit mean square there are no misfitting categories, though Category 1 was not used at all and Category 3 was predominantly used in the present test evaluation. This is also graphically presented in Figure 1.

Table 1 Category Function

	CATEGORY LABEL	OBSERVED SCORE	COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	STRUCTURE MEASURE	CATEGORY MEASURE	
Fair	2	2	17	4	8.4	-15.3	1.25	1.21	NONE	(-40.20)	2
Good	3	3	261	62	8.8	9.7	.98	1.18	-29.19	.00	3
Very Good	4	4	125	0	36.6	35.6	.88	.85	29.19	(40.20)	4
	MISSING		7	4	-6.9						

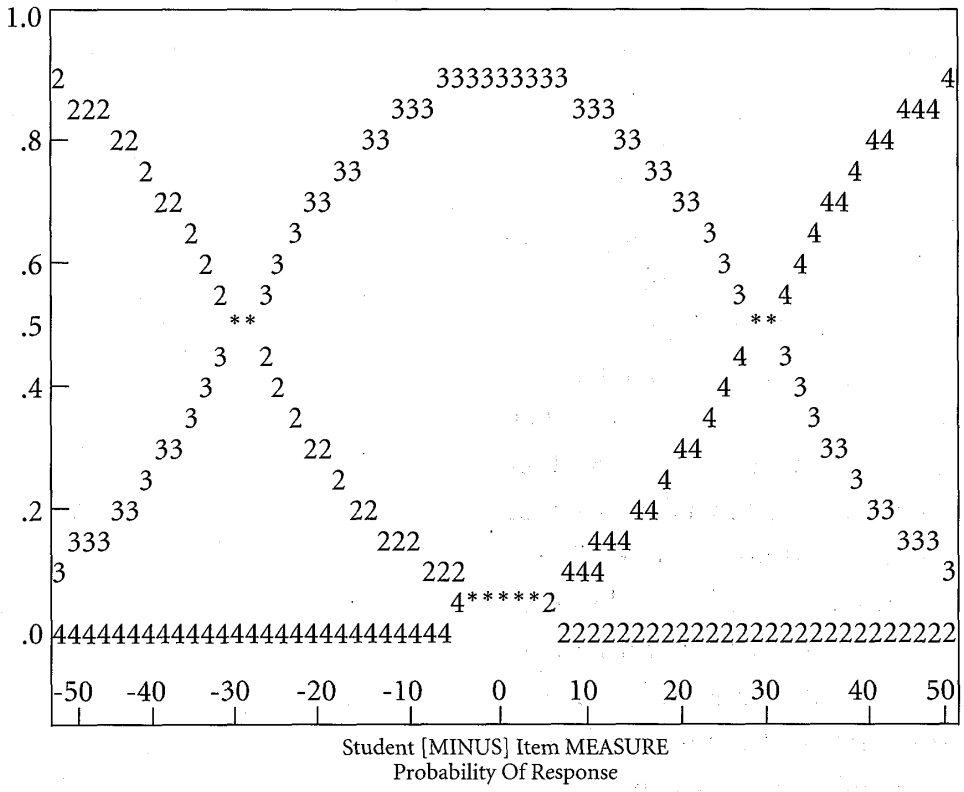

Figure 1 Category Probabilities: Modes

Figure 2 shows relative positions between two facets (students' ability and item difficulty) in a wider perspective. In the students' column, the top three students are the most able students, while only one least able student is at the bottom.

In the item column, Multilogue Grammar is the most difficult item, whereas Dialogue Conversation Strategies is the easiest one.

If we look at the distributions between students and items, it can be said that we need more difficult items to match better students.

Let us pay closer attention to the construct of items in the order of difficulty in this table. In the dialogue test, (items indicated by D), Grammar is the most difficult item followed by fluency and vocabulary, while conversations strategies is the easiest. Sound System (in other words: Pronunciation) is in the middle of the difficulty order.

In the multilogue test, (items indicated by M), again Grammar is the most difficult item followed by vocabulary and fluency, whereas conversation strategies and content are rather easy.

It may be that for both Dialogue and Multilogue Tests, grammar, vocabulary and fluency are in one large difficult group.

It is interesting to note that Dialogue Conversation Strategies is much easier than Multilogue Conversation Strategies. One possible explanation is that in a dialogue type face-to-face situation, students can easily use some appropriate phrases such as, "I beg your pardon," or "Could you say it again please?" as a strategy, whereas in a multilogue type discussion situation students have difficulty in using timely and appropriate expressions to interrupt and get involved in the talk.

Anther interesting thing is that the rater tends to be harsh especially on grammar. One reason for this is that it is easier to find ungrammatical and inappropriate sentences in students' spoken utterances.

Still, another interesting thing is that teachers are not fastidious about the sound problem (pronunciation) as long as the students are audible and comprehensible

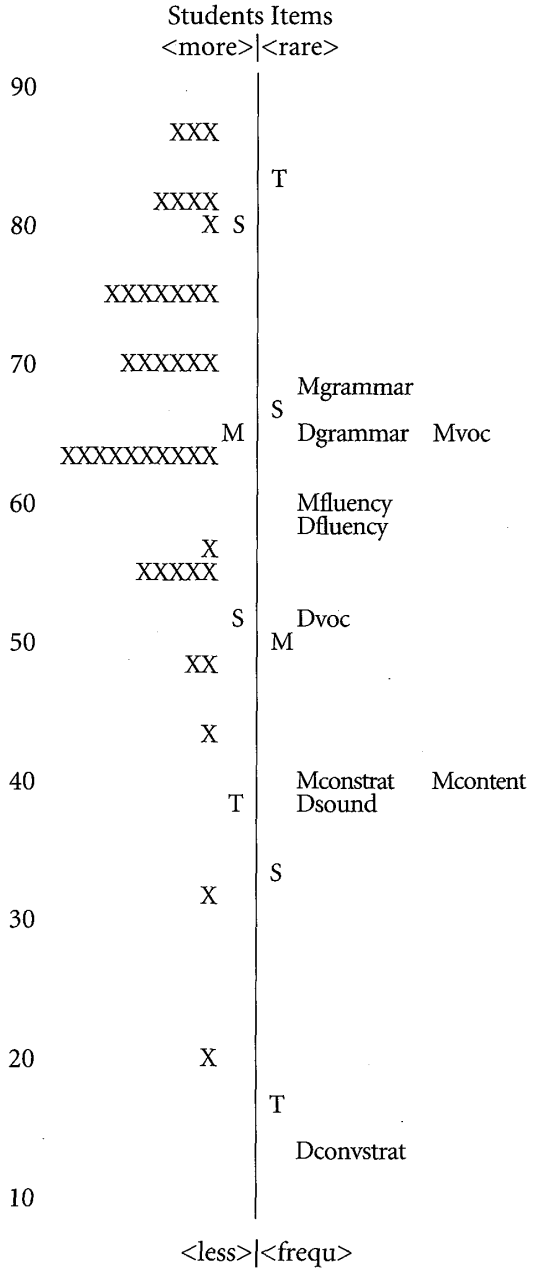

Figure 2 Item Map

Table 2 indicates the mean difference between the Dialogue (D) and Multilogue (M) Tests in terms of item difficulty showing that the first one as a whole is easier than the second one. This is understandable when we consider the complexity of the Multilouge situation.

Figure 3 shows expected measures. Mid-evel students get 2 or 3 points and poor side students in the left hand corner mainly get 2 points, while good students in the right hand corner mainly get 3 points and 4 points. This table also tells us the difficulty order of the items-Multilogue Grammar is the most difficult whereas Dialogue Conversation Strategies is the easiest, as we have seen in the previous data result.

Table 2 Comparison of Means between Dialogue Test and Multilogue Test Results

Item COUNT	MEAN MEASURE	S.E. MEAN	OBSERVED S.D.	MEDIAN	REAL SEPARATION	CODE
10	50.00	5.43	16.28	55.12	3.55	*
5	45.22	9.18	18.37	52.18	3.83	Dia
5	54.78	6.06	12.11	60.48	2.72	Multi

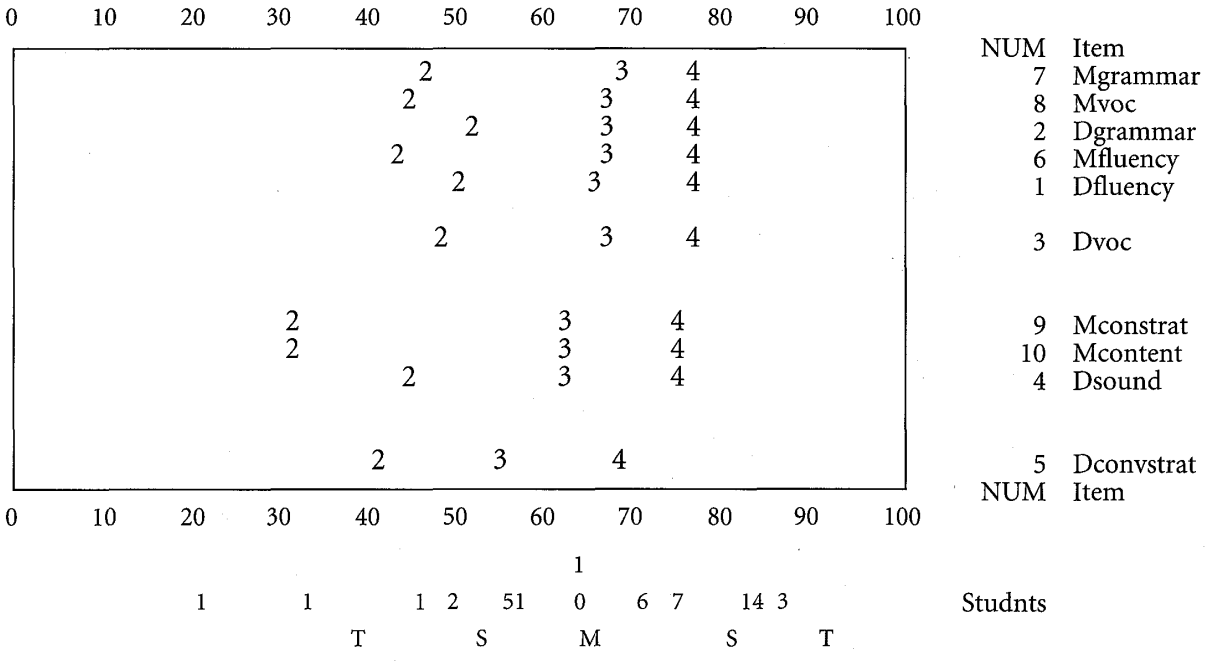

Figure 3 Expected Average Mesures by Category Score for Students

Table 3 presents 3 items (Multilogue vocabulary, Multilogue grammar, and Dialogue conversation strategies) which are weak in score correlation. However, none of them are negative, which means that all the 10 items including these three weak ones are going at least in the same direction. Thus, they may not cause any serious problems.

From the viewpoint of fit statistics we might want to look into Dialogue conversation strategies, because the outfit statistic Mean Square (1.70) is rather high and it is worthy of examination.

Table 4, in the column of outfit Mean Square, we can see some evidence that could prove the reason for Item Dialouge conversation strategies being misfit.

Table 3 Items Statistics: Correlation Order

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	REALSE	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	Items
8	117	39	65.4	4.3	.33	-2.8	.26	-2.8	.00	Mvoc
7	119	40	67.6	4.3	.50	-1.8	.51	-1.6	.05	Mgrammar
5	159	41	12.9	6.4	1.48	1.2	1.70	.8	.06	Dconvstrat
3	132	41	52.2	3.7	.96	-.2	1.01	.0	.26	Dvoc
2	124	41	64.7	4.2	.54	-1.8	.49	-1.7	.30	Dgrammar
1	128	41	58.1	4.3	1.16	.6	1.36	1.0	.34	Dfluency
4	143	41	38.3	4.2	1.44	2.4	1.50	2.1	.41	Dsound
6	123	40	60.5	5.0	1.49	1.5	1.50	1.2	.74	Mfluency
9	134	39	40.4	3.5	.87	-.8	.89	-.6	.78	Mconstrat
10	138	40	40.1	3.5	.93	-.4	1.03	.2	.83	Mcontent
MEAN	132.	40.	50.0	4.3	.97	-.2	1.02	-.1		
S.D.	12.	1.	16.3	.8	.40	1.6	.47	1.5		

Table 4 Items Category/Option/Distractor Frequencies: Misfit Order

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE MEASURE	S.E MEAN	OUTF MNSQ	Item	
5 A	3	3	5	12	65.00	3.37	1.7	Dconvstrat	3 Good
	4	4	36	88	67.07	2.16	1.5		4 Very Good
	MISSING***		5	12	20.26				

Let us go further to Table 5, which also can provide us the reason for Dialouge conversation strategies from the viewpoint of most unexpected response patterns. 4 intermediate and advanced level students were given 3 points in spite of their rather high abilities against an easy item Dialouge conversation strategies, whose measure is 12.9: by far the easiest one of all the ten items. 3 points given to them in Table 5 indicate the unexpected responses, which means the students are supposed to be given higher points (in this case, 4 points) when we consider their ability and the easiness of the item.

Table 5 Most Unexpected Responses

Item	MEASURE		Studnt
			22 3224211422 4332
			575131064054276420
		high	
4	Dsound	38.3 B	. . 3. . 3. . 33. . . . .
5	Dconvstrat	12.9 A	3. . . . . 2. . . . . 2. .
10	Mcontent	40.1 E	. . . . . 22.
9	Mconstrat	40.4 d	. . . . . 2.
3	Dvoc	52.2 e	. . . . . 44. . . 4. . .
1	Dfluency	58.1 D	. . . . . 4. . 442. . 4. . .
6	Mfluency	60.5 C	. 442. . . . . 222. . . .
2	Dgrammar	64.7 c	. . . . 4. . . . .
8	Mvoc	65.4 a	 3
7	Mgrammar	67.6 b	. . . . . 2. 3
			low
			225322421142274332
			57 1310640542 6420

Table 6 suggests further concrete evidence. Student No. 5 whose measure is (75.3) was given 3 points when he or she was expected to have 4 points in the item Dialogue conversation strategies. Student No. 21, whose measure is (69.2) was given 3 points when he or she was expected to have 4 points. Student No. 10, whose measure is (62.5) was given 3 points when he or she was expected to have 4 points. The same is true for Student No. 14.

In this way, we can examine the cause of the misfitting item and eventually improve the test by asking the raters about their rating procedure and by asking students about their performance.

Tables 7, 8 and 9 propose the results of factor analysis. (we take items with factor loading over .50)

Table 7 shows Factor 1. This factor can be called Multilogue Ability, though two types of components (one type: Multilogue Vocabulary and Multilogue Grammar, and the other: Multilogue Conversation Strategies and Multilouge Content) mainly contribute to Factor 1 in completely opposite directions. In other words, there is an important element in this factor which distinguishes between Vocabulary-Grammar and Content Conversation Strategies. Furthermore, Multilogue (Vocabulary and Grammar) are different from Multilogue(Content and Conversation Strategies).

Table 8 demonstrates Factor 2. It can be named Dialogue Ability, although two types of Dialogue items, fluency-vocabulary and : conversation strategies, mainly contribute to Factor 2 in completely opposite directions. This indicates that there is an important element which distinguishes between Dialogue-fluency and Dialogue-vocabulary and Dialogue Conversation strategies.

Table 9 illustrates Factor 3, Basic Sound System Handling Ability (Pronunciation), which is a substantial part of speaking.

Figure 4 presents a comparison of the students' measure scores on the Multilogue

test and Dialogue test. If we set a benchmark (.50) as a cut-off score between better students and poor students, we will roughly have two types of students: One consisting of students who are both good at Dialogue and Multilogue tests, and the other consisting of students who are good at Dialogue test but poor at Multilogue test.

Table 6 KEY: .1.= OBSERVED, 1 = EXPECTED, (1) = OBSERVED, BUT VERY UNEXPECTED.

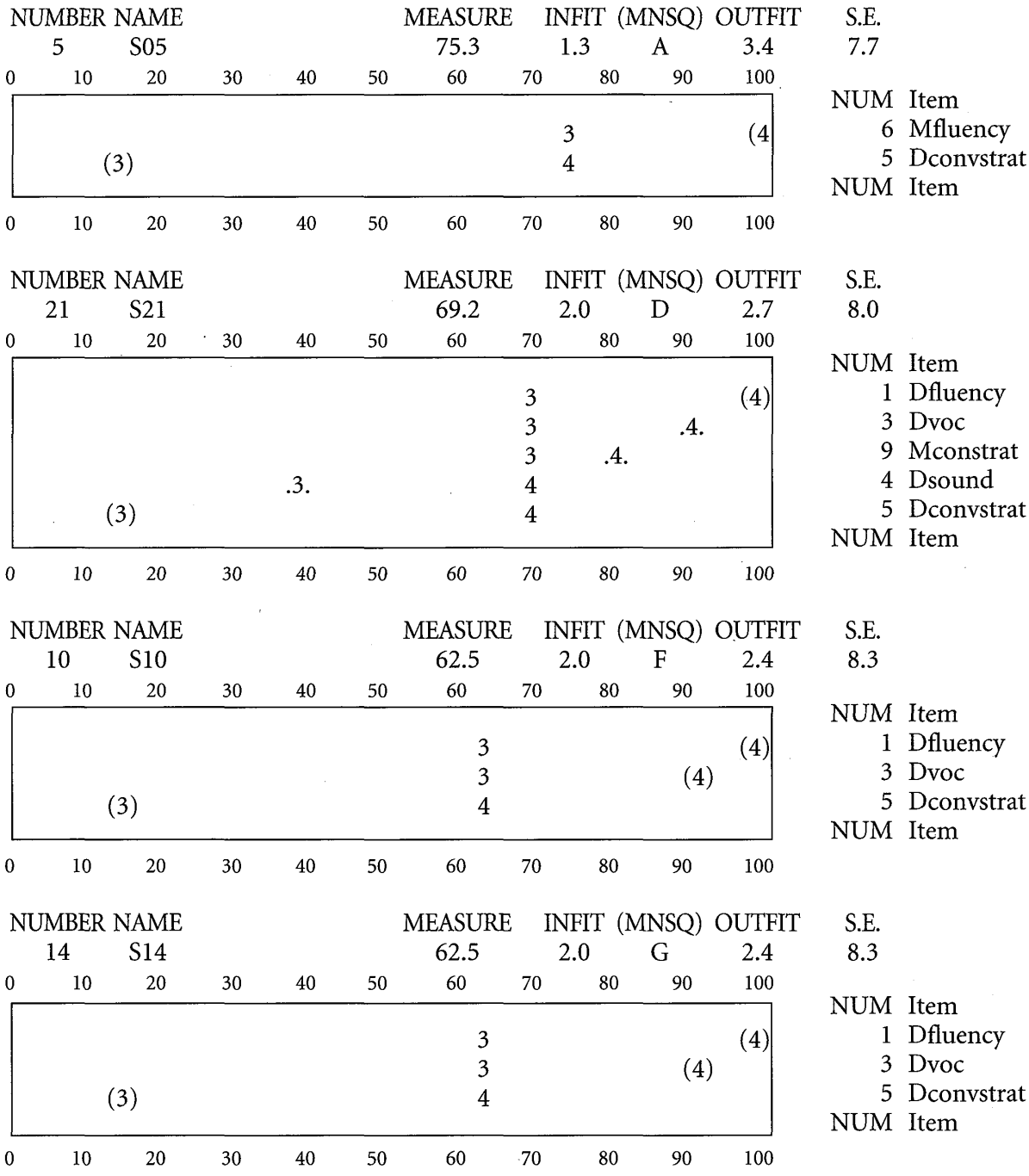

Table 7 FACTOR 1

FACTOR	LOADING	MEASURE	INFIT MNSQ	OUTFIT MNSQ	ENTRY NUMBER		Item
1	.84	65.4	.33	.26	A	8	Mvoc
1	.75	67.6	.50	.51	B	7	Mgrammar
1	.49	52.2	.96	1.01	C	3	Dvoc
1	.46	64.7	.54	.49	D	2	Dgrammar
1	.35	58.1	1.16	1.36	E	1	Dfluency
1	.01	38.3	1.44	1.50	e	4	Dsoun
1	.00	12.9	1.48	1.70	d	5	Dconvstrat
1	-.73	40.1	.93	1.03	a	10	Mcontent
1	-.71	40.4	.87	.89	b	9	Mconstrat
1	-.55	60.5	1.49	1.50	c	6	Mfluency

Table 8 FACTOR 2

FACTOR	LOADING	MEASURE	INFIT MNSQ	OUTFIT MNSQ	ENTRY NUMBER		Item
2	.82	58.1	1.16	1.36	E	1	Dfluency
2	.68	52.2	.96	1.01	C	3	Dvoc
2	.13	40.4	.87	.89	b	9	Mconstrat
2	.11	60.5	1.49	1.50	c	6	Mfluency
2	.04	40.1	.93	1.03	a	10	Mcontent
2	-.79	12.9	1.48	1.70	d	5	Dconvstrat
2	-.42	38.3	1.44	1.50	e	4	Dsound
2	-.26	65.4	.33	.26	A	8	Mvoc
2	-.19	67.6	.50	.51	B	7	Mgrammar
2	-.14	64.7	.54	.49	D	2	Dgrammar

Table 9 FACTOR 3

FACTOR	LOADING	MEASURE	INFIT MNSQ	OUTFIT MNSQ	ENTRY NUMBER		Item
3	.82	38.3	1.44	1.50	e	4	Dsound
3	.25	52.2	.96	1.01	C	3	Dvoc
3	.04	40.1	.93	1.03	a	10	Mcontent
3	-.55	64.7	.54	.49	D	2	Dgrammar
3	-.33	60.5	1.49	1.50	c	6	Mfluency
3	-.28	40.4	.87	.89	b	9	Mconstrat
3	-.15	65.4	.33	.26	A	8	Mvoc
3	-.14	67.6	.50	.51	B	7	Mgrammar
3	-.13	12.9	1.48	1.70	d	5	Dconvstrat
3	.00	58.1	1.16	1.36	E	1	Dfluency

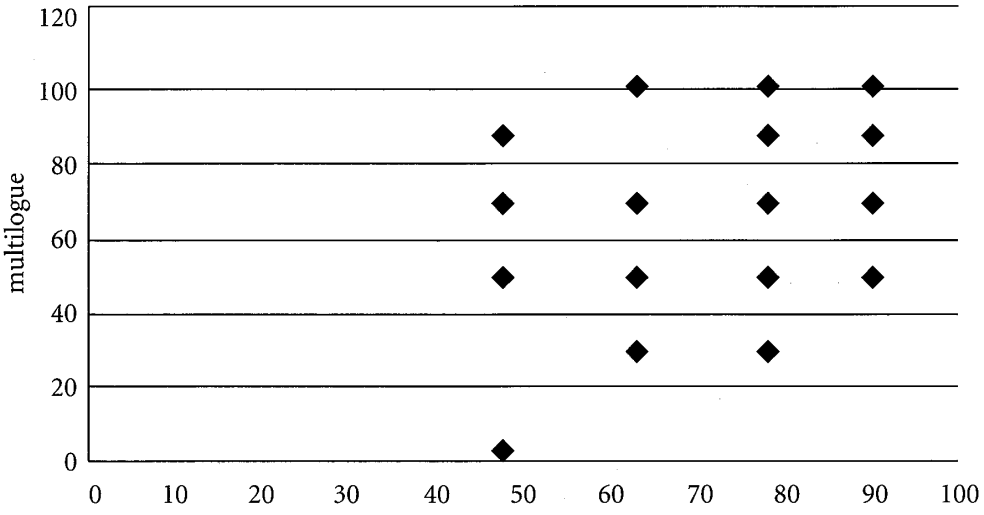

Figure 4 Comparison of Diallouge and Multilogue (persons)

N. B. Some students obtained the same scores, so there are some cases where more than two students overlap on the same spot.

5. Conclusions and Implications

We can draw the following conclusions:

1. Through the item map where relative positions of students and items are shown, we need more difficult items in order to match better students.
2. As a whole, the multilogue test is more difficult than the dialogue test, athough in each test, the difficulty order of each item is slightly different.
3. Among the ten evaluation items, Dialogue Conversation Strategies is the easi est. Furthermore, this is easier than Multilogue Conversation Strategies. It may be that conversation strategies can be more easily used in the Dialogue setting than in the Multilogue setting.
4. Teachers are generally severe about grammar, while they are rather lenient about pronunciation in the speaking test situation.
5. Ten evaluation items are measuring the speaking ability in the same direction, although the degree of contribution varies from item to item.
6. Through the investigation of unexpect ed response patterns of misfitting items in terms of students' observed points, we were able to find the interaction among the rater, the students and the items, and eventually this result can be used for the improvement of the test, such as for the rater training and the item rearrangement.
7. Factor analysis shows that speaking abil ity is composed of at least three compo nents (Monologue, Dialogue and Sound), although other possible elements can be added judging from the complexity of spoken utterances.
8. Students can be categorised into two groups (one which is good at both Multi and Dialogue tests, the other which is only good at Dialouge test)
9. One suggested implication for the class room is to enhance students' Multilougue speaking ability by providing appropriate learning situations in class room settings.

Bibliography

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition* 10,149-64.

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

Bachman, L. F., Lynch, B. and Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing* 12, 238-56.

Bond , T.G. and Fox, C.M.(2001), *Applying the Rasch Model: Fundamental Measurement in the Human Sciences.* Lawrence Erlbaum Associates: Mahwah, New Jersey.

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

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing* 20,1,1-25.

Fulcher, G. (1996). Testing tasks; issues in task design and the group oral. *Language Testing* 13, 23-51.

Linacre, J. M. (1997). Guidelines for rating scales. *Mesa Research Note 2* (Online). Available at http://www.rasch.org/rn2.htm.

Linacre, J. M. (1998a). FACETS 3.17. Computer program. Chicago, Ill MESA Press.

Linacre, J. M. (1998b). Rasch first or factor first? *Rasch Measurement Transactions* 11, 603.

Linacre, J.M. (1999a). *A user's guide to Facets; Rasch measurement computer program.* Chicago, Ill: MESA Press.

Linacre, J. M. (1999b). How muc is enough? *Rasch Measurement Transactions* 12, 653.

Lynch, B. K. and McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing* 15, 158-80.

McNamara, T (1996). *Measuring second language performance.* Harlow: Addison Wesley Longman.

McNamara, T.F. and Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing* 14, 140-56.

Upshur, J. and Turner, C. (1999). Systematic effects in the rating of second-language speaking ability: test method and learner discourse. *Language Testing* 16, 82-508.

Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing* 15, 263-87.

Acknowlegement:

I am grateful to Dr. Garold Murray for his cordial help in the process of this project.

Appendix

Criteria for Dialogue Test	1	2	3	4
Fluency	Single words; long pauses	Conversation not smooth; frequent pauses; ideas may be disconnected	Conversation somewhat smooth; some pauses, topic shift not smooth	Conversation smooth; few pauses; topic shift OK
Grammar	Single words; Many errors; Meaning not clear	Limited patterns; many errors; meaning may not be clear	Occasional errors may make meaning unclear	Some errors; meaning clear; complex syntax
Vocabulary	Very limited; may use Japanese	Poor; many pauses; can express basic ideas; a Japanese word or two	Fair; searches for words; may make errors	Good; few pauses while searching for words
Conversation Strategies	No conversation; single words; doesn't repair breakdowns	Often inappropriate: tried to have a conversation; use of L1 to repair breakdowns	Occasionally inappropriate; some difficulty repairing breakdowns	Generally appropriate; repairs breakdowns
Sound System	Difficult to understand	Meaning sometimes not clear; using Katakana sounds	Occasional errors; meaning clear	Some errors; meaning clear

Criteria for Multilogue Test	1	2	3	4
Fluency	Single words; long pauses	Conversation not smooth; frequent pauses; ideas may be disconnected	Conversation somewhat smooth; some pauses, topic shift not smooth	Conversation smooth; few pauses; topic shift OK
Grammar	Single words; Many errors; Meaning not clear	Limited patterns; many errors; meaning may not be clear	Occasional errors may make meaning unclear	Some errors; meaning clear; complex syntax
Vocabulary	Very limited; may use Japanese	Poor; many pauses; can express basic ideas; a Japanese word or two	Fair; searches for words; may make errors	Good; few pauses while searching for words
Conversation Strategies	No conversation; single words; doesn't repair breakdowns	May respond; may use L1 to repair breakdowns; no questions	Generally responds and agrees or disagrees; some difficulty repairing breakdowns	Responds; agrees or disagrees; asks questions; repairs breakdowns
Sound System	Very little said, no details	Makes one or two points	Expresses opinions some details	Expresses opinions freely; supports ideas