ラッシュモデルを用いた口頭発表能力テストのデータ分析
Rasch Based Analysis of Oral Proficiency Test Data

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

Keywords 口頭発表能力テスト，ラッシュモデル，項目図，因子分析
Oral Proficiency Test, Rasch Model, Item Map, Factor Analysis

ABSTRACT

本稿は英語口頭発表能力テストのデータ分析を項目応答理論の 1 つであるラッシュモデルを用いて，テスト項目の分布図および因子分析により口頭発表能力の構造的枠組をより科学的に行おうとするものである．中村 (1999) の 2 因子提案をもとに，今回の研究では項目応答理論の基本概念である unidimensionality, および local independence の観点から分析を加え，その結果 3 因子構造が確認された．結論として Monologue, Dialogue, Multilogue の 3 つの能力要因が独立的ではあるが，全体として Oral Proficiency という大きな枠組みの 1 つの能力を測定していると考えられる．

Rasch Based Analysis of Oral Proficiency Test Data

This paper examines the rating scale data of Oral Proficiency Tests, which has already been analyzed through raw scores, from the viewpoint of the Rasch analysis focusing on two things: 1) item map and 2) factor analysis.

First, we will discuss the item map. Nakamura (1999) argues about the difficulty order of 6 items and the students' answering patterns using descriptive statistics and the distributions of frequency of students' test scores. The data shows that TMS is the easiest item followed by GI, SO, SGD, FFI, and LGD, which is the most difficult one. The data also shows that the lowest point is rarely used except in LGD, which automatically tells us that LGD is the difficult item. All the information is given without a device of item map.

N. B. Abbreviation Used:
SO: Speech Making Overall
TMS: Tape Mediated Sociolinguistic Test
FFI: Face to Face Interview
GI: Group Interview
SGD: Small Group Discussion Test
LGD: Large Group Discussion Test

Table 1 is an item map prepared by the Rasch measured scores in the present research. Table 1 shows us another way of looking at the relationship between person and item. The figure shows the position of each of the four rating categories (very good: 4, good: 3, fair: 2, poor: 1) for the six items arranged in order of difficulty. This type of map can be used, as Bode and Wright (1999) claim, to develop a quick-scoring method that takes the difficulty level of individual categories and items into account. Here, we can observe the fact that, in LGD, students were rated in a good balance by being given the lowest point to a certain number, while in GI and TMS, very few students were given the lowest point. In other words, LGD turned out to be the most difficult item. Compared with Nakamura (1999), this item map information can be quicker and more comprehensive with an easier data handling process.

Table 1

INPUT: 46 STUDENTS, 6 ITEMS ANALYZED 45 STUDENTS, 6 ITEMS 4 CATS WINSTEPS v2.85

```
EXPECTED SCORE: MEAN  (":" INDICATES HALF-SCORE POINT)
 0   10    20    30    40    50    60    70    80    90   100
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM   ITEM
1                 1    :      (2)    :       3        :      44    6  LGD
|                                                              |
1             1     :     (2)     :       3        :     4    4    3  FFI
|                                                              |
1          1    :       2     :       (3)       :     4       4    5  SGD
|                                                              |
1        1    :       2     :       (3)       :     4         4    1  SO
|                                                              |
1     1    :       2     :       (3)       :     4            4    4  GI
1     1    :     2       :       (3)       :     4            4    2  TMS
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM   ITEM
 0   10    20    30    40    (50)   60    70    80    90   100

                1      3 2  3 5 6  2  5 6   6  2   3     1   1  STUDENT
                  Q       S       M       S       Q
```

Bode and Wright (1999) state further that: "One can record item responses on such a map, determine the approximate average horizontal position by eye, and draw a vertical line down to the expected score at the bottom to estimate the overall measure. Unexpected responses which digress from the vertical line are easily spotted and can be used diagnostically. The empty spaces between items indicate a signifi-cant difference in their difficulty — a difference greater than two standard errors of their calib-ration estimates." (p. 309)

This type of item map can also be used to describe the frequency of activities reported by an individual student (cf. Bode and Wright 1999). Let us look at an example of the location of category labels vertically above a measure of 50.

Consider a student with a measure of 50, who is expected to get 16 points in total from the 6 test items as indicated in the circled numbers in Table 1. This student is rated above the middle point (2.5 in the 1-4 scale) in TMS, GI, SO and SGD, while he / she is rated below the middle point in FFI and LGD. In GI and TMS, this student is rated closer to 3 points, whereas in LGD he / she is rated almost on 2 points. Overall, this average student is expected to get lower points than the middle point in LGD and FFI, which seems to explain the difficulty level of these items. However, if this student gets only one point in TMS, there is something wrong with this student. We should diagnostically investigate it immediately.

Let us take a look at another example. This time, consider a student with a measure of 80 (in Table 1-a). This student is expected to get 4 points in TMS, GI, SO and SGD, and 3 points in FFI and LGD. If this student gets only 2 points in LGD, something is wrong with this student or the item. We should diagnostically examine the reason which has caused the misfitting case. Thus , this item map can help to locate the misfitting part quite quickly.

Table 1-a

INPUT: 46 STUDENTS, 6 ITEMS ANALYZED 45 STUDENTS, 6 ITEMS 4 CATS WINSTEPS v2.85

```
EXPECTED SCORE: MEAN  (":" INDICATES HALF-SCORE POINT)
 0   10    20    30    40    50    60    70    80    90   100
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM   ITEM
1                1    :     2     :      (3)      :      44    6  LGD
|                                                            |
1             1     :     2     :      (3)       :     4   4    3  FFI
|                                                            |
1          1    :      2     :       3       :     (4)      4    5  SGD
|                                                            |
1        1    :      2     :       3       :     (4)        4    1  SO
|                                                            |
1     1    :      2     :       3       :     (4)           4    4  GI
1     1    :     2      :      3        :     (4)           4    2  TMS
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM   ITEM
 0   10    20    30    40    50    60    70   (80)   90   100

               1      3 2  3 5 6  2  5 6   6  2   3     1   1  STUDENT
                  Q       S       M       S       Q
```

Table 1-b

INPUT: 46 STUDENTS, 6 ITEMS ANALYZED 45 STUDENTS, 6 ITEMS 4 CATS WINSTEPS v2.85

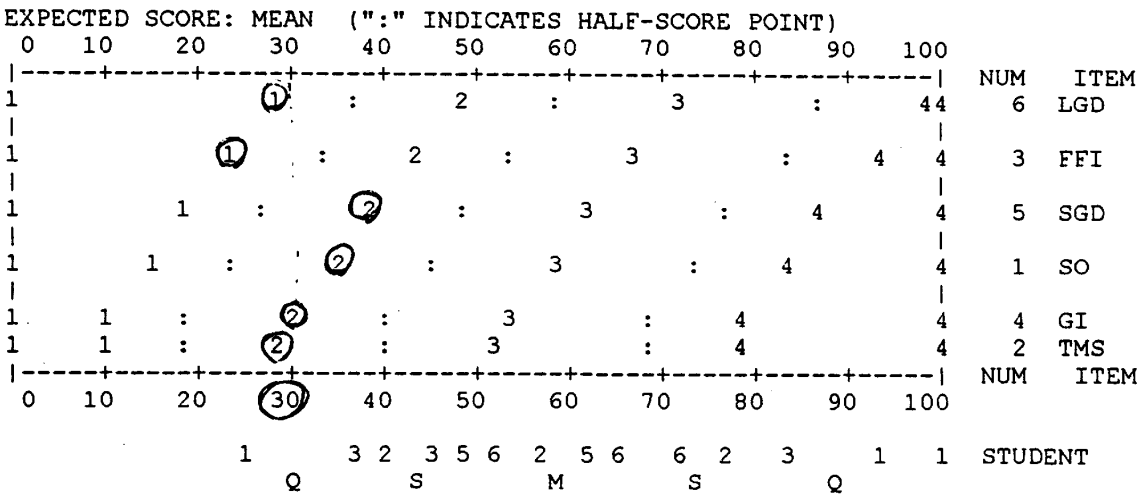

Table 2

FACTOR 1 FROM PRINCIPAL COMPONENT ANALYSIS OF STANDARDIZED RESIDUAL COR-RELATIONS FOR ITEMS (SORTED BY LOADING) FACTOR 1 EXPLAINS 1.83 OF 6 VARIANCE UNITS

FACTOR	LOADING	MEASURE	INFIT MNSQ	OUTFIT MNSQ	ENTRY NUMBER	ITEM
1	.78	41.8	.91	1.01	B	2 TMS
1	.74	47.1	.83	.80	c	1 SO
1	.02	56.6	.70	.71	b	3FFI
1	-.59	50.8	.66	.66	a	5 SGD
1	-.55	61.0	1.71	1.68	A	6 LGD
1	-.13	42.5	.97	.98	C	4 GI

Let us look at yet another example. This time, consider a student with a measure of 30 (in Table 1-b). This student is expected to obtain 10 points in total. If this student gets 4 points in FFI, there is something wrong with this student or the item. We should diagnostically analyze the reason for this misfitting case.

Secondly, we will talk about the factor analysis. Nakamura (1999) states that two factors were obtained through a factor analysis using raw scores from the test results. He suggested that one factor should include the functioning of the number of people involved in the oral language activities, while the other is related to a linguistic element.

The present research will investigate the factors from another viewpoint by employing the Rasch analysis and explore the details of the factors.

Tables 2, 3, and 4 below show that we are able to obtain 3 factors. The first factor is composed of two tests: TMS and SO. The second factor consists of LGD, and the third

factor is contributed by GI. The rest of the items (tests) were not involved in the 3 main factors; however, SGD and LGD strongly show their opposite contribution to the first factor, which indicates that the number of people in the test is important. Furthermore, FFI demonstrates an extremely strong reverse contribution to the second factor, which also suggests the significance in the number of people involved

Table 2 shows Factor 1. This factor can be called Monologue Ability because in SO and TMS, students are speaking to the tape by themselves in the language laboratory, even though there is a semi-direct interaction between a student and the stimulus (which is heard from the recorded tape). As mentioned above, SGD and LGD are making contributions to Factor 1 in the opposite directions. This indicates that there is an important element in this factor (Monologue Ability), which distinguishes between Monologue Ability and Non-monologue ability. Furthermore, this Monologue Ability suggests that one's Monologue Ability is different from one's Non-monologue ability.

Table 3 demonstrates the second factor. This factor, Factor 2, is made only of LGD, and can be named Multilogue Ability, because in LGD a student needs to demonstrate his / her discussion ability in a large class sized group (more than 10 people involved). Though TMS is included, it can be ignored as a contributing element to this factor, due to the small factor loading (below .30). Actually, TMS has already contributed to Factor 1 (Monologue Ability).

What should be noticed in Factor 2 is that FFI and SGD (especially FFI) are contributing to this factor in the opposite directions. This indicates that there is an important element which is distinguishing between LGD and FFI and SGD, which could be due to the number of people involved.

Table 4 shows the third factor, Factor 3. This factor is supported by GI, and can be called Dialogue Ability because a student should respond to questions asked by an interviewer (interviewers), even though the situation is not face to face nor one on one.

Although Nakamura (1999) started with six different tests to evaluate students' communicative language ability, and ended up with two factors (a person related factor and a linguistic factor), the result of the Rasch based analysis

Table 3

FACTOR 2 FROM PRINCIPAL COMPONENT ANALYSIS OF STANDARDIZED RESIDUAL COR-RELATIONS FOR ITEMS (SORTED BY LOADING) FACTOR 2 EXPLAINS 1.57 OF 6 VARIANCE UNITS

FACTOR	LOADING	MEASURE	INFIT MNSQ	OUTFIT MNSQ	ENTRY NUMBER	ITEM
2	.79	60.1	1.71	1.68	A	6 LGD
2	.26	41.8	.91	1.01	B	2 TMS
2	-.76	56.6	.70	.71	b	3 FFI ·
2	-.49	50.8	.66	.66	a	5 SGD
2	-.24	42.5	.97	.98	C	4 GI
2	-.09	47.1	.83	.80	c	1 SO

Table 4

FACTOR 3 FROM PRINCIPAL COMPONENT ANALYSIS OF STANDARDIZED RESIDUAL COR-RELATIONS FOR ITEMS (SORTED BY LOADING) FACTOR 3 EXPLAINS 1.14 OF 6 VARIANCE UNITS

FACTOR	LOADING	MEASURE	INFIT MNSQ	OUTFIT MNSQ	ENTRY NUMBER	ITEM
3	.96	42.5	.97	.98	C	4 GI
3	-.34	56.6	.70	.71	b	3 FFI
3	-.26	50.8	.66	.66	a	5 SGD
3	-.16	61.0	1.71	1.68	A	6 LGD
3	-.12	41.8	.91	1.01	B	2 TMS
3	-.01	47.1	.83	.80	c	1 SO

suggests that we need three tests to make a more precise measurement of students' communicative language ability: the first is a test for Monologue Ability, the second is a test for Multilogue Ability, and the third one is a test for Dialogue Ability.

Putting it another way, we tend to think that one test will give enough information to understand students' speaking ability usually from practical reasons. However, the present Rasch analysis result indicates that we need to look at their communication ability from multidimensional viewpoints such as Monologue, Dialogue and Multilogue, so that we can conduct a more accurate measurement.

Through TMS or SO tests we can measure Monologue Ability in which students express their basic speaking ability on tape. By using a GI test, we can assess Dialogue Ability in which students interact with the live interviewer in a small group. Through the method of a LGD test, though it is a unique aspect of speaking ability, we can measure Multilogue Ability where students perform using their ability of argumentation, discussion and debating.

Granted, the ideal theoretical construct through the Rasch analysis should be more or less unidimensional, we were, in practice, able to obtain three factors that give us different views or angles at which to look at communication ability. In other words, we can view the communicative language ability as multidimensional (Monologue, Dialogue and Multilogue) from the practical viewpoints at each individual stage, though each dimension makes a great contribution on its own to construct the whole unidimensional communicative language ability.

In conclusion, we have analyzed Oral Proficiency Tests Data using the Rasch measured scores by focusing on two things: 1) item map, and) factor analysis. We have been able to utilize the idea of an item map in order to spot the level of difficulty of items, and general view of students' expected responses.

The results of the factor analysis have provided information for the existence of three factors, which in practice are necessary to measure students' Communicative Language Ability more accurately. However, as a whole construct of the language ability, the unidimensional view of Communicative Language Ability has also been proposed.

Acknowledgement

I am grateful to Dr. Benjamin D. Wright and Dr. John M. Linacre for their invaluable comments.

Bibliography

Bode, R. K., and 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.

Linacre, J. M. (1989, 1993, 1994). *Many-Facet: Rasch Measurement.* Chicago: MESA Press.

Nakamura, Y. (1999). Measuring speaking skills through multidimensional perfor-mance tests. *Educational Studies*, 41 99-113. Tokyo: International Christian University.

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

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 S. Emretson, and S. Hershberger (Eds.). *The New Roles of Measurement: What Every Psychologist and Educator Should Know* (pp. 65-104). Hillsdale NJ: Lawrence Erlbaum Associates.

Wright, B. D. and Linacre, J. M. (1998). *A User's Guide to WINSTEPS / BIGSTEPS: Rasch-Model Computer Programs.* Chicago: MESA Press.

Wright, B. D., and Masters, G. N. (1982). *Rating Scale Analysis: Rasch Measurement.* Chicago: MESA Press.

Wright, B. D, and Stone, M. H. (1979). *Best Test Design: Rasch Measure-ment. Chicago:* MESA Press.