ファセット・ラッシュモデルによる 英作文テストデータの分析

Writing Test Data Analysis with a Many-Faceted Rasch Model

中村　優治 NAKAMURA, Yuji

● 東京経済大学
Tokyo Keizai University

Keywords
ファセット・モデル，英作文テスト，評価者の容認度，データ分析
FACETED Model, Writing Test, Rater's Leniency, Data Analysis

ABSTRACT

本稿は英作文テストデータの分析を FACETS 理論に基づいて行い，この理論の実用性を 3 つの側面（学生の作文能力，評価者の容認度，評価項目の機能）から科学的，統計的に検証しようと試みたものである．結論として，この分析を通して，これまで古典的テスト理論では説明が困難であった同一スケール上での上記 3 つの側面の比較が可能になったこと，また，評価項目あるいはテストタスクの観点からテストの改善への手がかりが得られるようになったことなどがあげられる．

1. Theoretical background and rationale

Application of Rasch Measurement Models to production testing has been one of the issues in language testing. This paper will focus on the value of using Rasch analysis in describing and assessing students' performance in writing, as well as the raters' severity. The eventual goal of this study is to find whether the test items define a meaningful ability variable, the consistency of the raters in their grading and if the marking scheme used needs to be further modified.

2. Research design and methods

15 Japanese college students took a writing test consisting of two tasks (a narrative story and a conversational story based on the cartoons they chose). The composition data were scored in terms of eight evaluation items such as grammar and discourse using a four-point scale by two teachers. The graded responses were analyzed using a many-faceted Rasch measurement (FACETS) model.

Task : Students wrote a story and a conversation based on cartoons
Raters : 2 raters (using 1-4 scale)
Items : 8 evaluation items
Grammar, Sociolinguistic competence, Interactional competence, Vocabulary, Overall, Discourse, Fluency, Content
Subjects : 15 Japanese university students
Acceptable range for the Infit and Outfit
Statistic : 0.6-1.4

3. Purpose of the research

This paper attempts to answer the following questions.

1) What does the Many-Facet Rasch Measurement tell us:
   (1) the function of rating categories?
   (2) the relationship among the three facets (students, items, raters)?
   (3) the students' ability?
   (4) the item difficulty (and or discrimination)?
   (5) the rater severity/leniency?
   (6) the function of tasks
2) How the assessment or the test can be improved by utilizing the analyzed data?

4. Results and discussion

The location of both person ability and item difficulty shows a spread of about 6 logits. Since the Rasch model deals with an assumed single underlying trait along which both items and persons are located on the same continuum, we can analyze the person's ability and item difficulty on the same scale

Table 1. Category Statistics

Model = ?,?,?,R

DATA				QUALITY CONTROL			STEP CALIBRATIONS		EXPECTATION		MOST PROBABLE	THURSTONE THRESHOLD	Cat PEAK
Category Score	Counts Used	%	Cum. %	Avge Meas	Exp. Meas	OUTFIT MnSq	Measure	S.E.	Measure at Category	at -0.5	from	at	Prob
1	35	16%	16%	-2.99	-3.06	1.1			(-3.89)		low	low	100%
2	86	38%	54%	-.96	-.84	.9	-2.80	.25	-1.18	-2.87	-2.80	-2.83	71%
3	62	28%	82%	1.11	1.08	.7	.46	.20	1.41	.27	.46	.36	56%
4	41	18%	100%	2.88	2.74	.8	2.34	.23	(3.52)	2.63	2.34	2.45	100%
									(Mean)		(Modal)	(Median)	

unidimensionally. Furthermore, in production tests such as writing tests, the raters' severity or harshness can also be examined by calculation of logits.

1) The function of rating categories

In Table 1, the column of outfit mean square shows the overall use of four categories (1=poor, 2, 3, 4=good), and they all function well within the acceptable range (0.6-1.4). This also means that the two raters use four categories in a reasonable way.

2) The relationship among the three facets (students, items, raters)

Table 2 shows relative positions among three facets (students' ability, item difficulty, and rater severity) in a wider perspective. In the student's column, Koda is the most able student while Aso is the least able one. In the item column "grammar" is the most difficult item whereas "content" is the easiest one. In the rater column, Rater A is more lenient than Rater B. Thus, we can have a quicker birds' eye view of the interrelations among the three facets in this "All facet vertical rulers" in Table 6. Let us pay special attention to the construct of items in the order of difficulty in this table. Grammar is the most difficult item followed by Sociolinguistic competence and Interactional competence, while Content is the easiest followed by Fluency and Discourse. Overall stands in the middle of the difficulty order. One possible explanation for the item Grammar is that in the written from it is easy for raters to find grammatical errors and naturally raters tend to be harsh on the students' writing. Also, in the written form it might be difficult for the students to express sociolinguistic competence and interactional competence, which could be easier in the spoken form in real oral communication settings. They may not be familiar with expressing conversational or soiciolinguistic phrases in the written form. The item all (overall), which is intended to grasp the general overview of the facets, tends to be the center of the difficulty. This is rather understandable.

The easier part is Content and Fluency. The students are good at creating original ideas about interpreting the cartoon, and therefore, their answers in other words can be creative. This can be influential to the raters' higher grades. Also, since students have only to follow the plot of the cartoons, the flow of ideas should naturally be smooth. The smooth development of the ideas can affect the raters which will be scored under the item (Fluency).

3) The students' ability

Table 3 demonstrates students' measurement report. The column of infit and outfit statistic indicates that student 4 and student 6 are overfitting, while, student 5 and student 10 are misfitting. Among them, student 5 is an extreme case of misfitting. Also, we can tell that Aso is the least able with the logit score of -3.50 while Koda is the most able student with the logit score of 2.28.

Now we will look into Student 5 more closely in Table 4 in order to find a possible explanation of his misfitting behavior. When we take a look at Table 4, in the column of outfit statistic, three categories (1, 2, 3) out of four are found to be misfit. This indicates that an inappropriate score is against his true ability of (-.12). This table shows that something is wrong with this student's behavior or product. In the columns of Average Measures and Expected Measures, the Average Measure

Table 2. All Facet Vertical "Rulers".

Measr	+ student	- item	- rater	S.1	S.2
3				(4)	(4)
				—	
	koda				
	kimu				
2					
					3
			B	3	
	ito seto	gra			
1					
	kume	soci			
	eto soto uda	intr			
		voc		—	—
0		all			
	oga	dis			
	shmi				
		flu			
	kenmo				
-1	sato sudo				
				2	
			A		
					2
		cont			
-2					
	kato				
				—	
-3					
	aso				
					—
-4				(1)	(1)
Measr	+ student	- item	- rater	S.1	S.2

N.B. : gra (Grammar), soci(Sociolinguistic comptence), intr(Interactional competence), voc(Vocabulary), all (Overall), dis (Discourse), flu(Fluency), cont(Content)

is expected to increase with the category value. In the present case, as the category goes up from 1 to 4, the expected measures go up from -1.99 to 2.16 However, in the actual average measure, there is disordering between the Average Measure in Category 1 (.14) and that in Category 2 (-.50). This means that this student is not measured appropriately according to his true ability. That is why in Table 3 student 5 (Oga) was picked up as an extreme case of misfitting.

Those students who turned out to be misfitting can be deleted from the statistical analysis, though they are to be investigated why they behave in that way. This will probably give us more information about what is happening internally in the misfitting students.

4) The item difficulty (and of discrimination)

Table 5 demonstrates that all the items except item 8 (all) are within the acceptable range of infit-outfit statistic, which is recognized in the column of infit and outfit statistic. Since this is the only misfitting item, it is worthwhile to consider deleting this item to see if there is any change. Let us look at Table 5. All the seven remaining items now look reasonable within the acceptable range. The item all (overall), which is an overfitting item as mentioned before, is to have a general ideal of students' performance. It tends to come to the center of the rating category. In other words, the score will be consistent, but no new or specific information is expected from this item. Coming toward the center is the nature of this item (overall), which is less informative and tends to have a misfitting value. However, we should not delete this item because it still gives us an overall view of students' ability.

5) The rater severity / leniency

Table 6 indicates that two raters behave quite reasonably within the acceptable range. Although their severities are quite opposite, each rates their students in a consistent way using the assigned rating scale and criteria, as shown in the column of Infit and Outfit statistic.

6) The function of tasks

By comparing the mean item measures of the two tasks (Task 1 and Task 2) in table 7, there is a statistically significant difference between the two at the .05 level of significance. In other words, Task 1 is easier than Task 2, which was statistically proved as below.

Items in Task 1 are composed of grammar, discourse, content, vocabulary and fluency, while those in Task 2 consist of interactional competence and sociolinguistic competence. Task 1 requires students to write a narrative story based on the cartoon they chose, whereas Task 2 requires students to write a conversation based on the cartoon. This proves that Task 1 (writing narrative stories) is easier for students that Task 2 (writing conversations). Putting it in another way, measuring students sociolinguistic and interactional competence through written forms is not an appropriate way. Probably more authentic performance tests such as interview tests or role play tests could be considered.

However, Table 8 and Table 9 indicate that items in Task 1 and items in Task 2 are functioning well, which was proved in the Infit and Outfit statistic column. Therefore, each individual item is working properly within each task. Still, a more appropriate method (again probably by conducting more authentic tests) could or should be given to more precisely measure students' performance ability.

5. Conclusions and implications

The following conclusions can be drawn. Firstly, the Rasch based analysis provides us with 1) the relationship among three facets (raters, students, items), 2) the rater severity and fit statistic, 3) the students' ability and fit statistic 4) the item difficulty and fit statistic, and 5) the functioning of rating categories. With all, or part of these pieces of information, the facets of the test can be thoroughly investigated individually, which was not possible in the traditional test analysis.

Secondly, the test can be improved by examining the fit statistic (misfit items) statistically not only in terms of individual items but also in terms of tasks.

Table 3. **Student Measurement Report** (arranged by N).

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu	student
23	16	1.4	1.34	-3.50	.56	1.1	0	1.3	0	1	aso
46	16	2.9	2.88	1.14	.43	1.0	0	.9	0	2	ito
43	16	2.7	2.64	.61	.42	.6	-1	.6	-1	3	uda
43	16	2.7	2.64	.61	.42	.6	-1	.5	-1	4	eto
39	16	2.4	2.39	-.12	.48	2.0	2	1.9	1	5	oga
26	16	1.6	1.57	-2.66	.51	.6	-1	.5	-1	6	kato
51	16	3.2	3.29	2.08	.45	.7	0	.6	0	7	kimu
44	16	2.8	2.72	.78	.42	.8	0	.9	0	8	kume
35	16	2.2	2.13	-.77	.42	1.1	0	1.1	0	9	kenmo
52	16	3.3	3.37	2.28	.45	1.6	1	1.4	0	10	koda
34	16	2.1	2.07	-.96	.43	.7	0	.7	0	11	sato
37	16	2.3	2.24	-.42	.42	1.4	1	1.2	0	12	shmi
34	16	2.1	2.07	-.96	.43	.8	0	.8	0	13	sudo
46	16	2.9	2.88	1.14	.43	.6	-1	.6	-1	14	seto
43	16	2.7	2.64	.61	.42	.9	0	.9	0	15	soto
39.7	16.0	2.5	2.46	-.01	.44	1.0	-.2	.9	-.3	Mean (Count : 15)	
8.0	.0	.5	.55	1.54	.04	.4	1.1	.4	.9	S.D.	

RMSE (Model) .45 Adj S.D. 1.48 Separation 3.32 Reliability .92
Fixed (all same) chi-square: 151.6 d.f.: 14 significance: .00
Random (normal) chi-square: 13.8 d.f.: 13 significance: .39

Table 4. **Student Oga's Data**

Model = 5,?,?,R oga

DATA				QUALITY CONTROL			STEP		EXPECTATION		MOST	THURSTONE	Cat
Category	Counts		Cum.	Avge	Exp.	OUTFIT	CALIBRATIONS		Measure at		PROBABLE	THRESHOLD	PEAK
Score	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	at	Prob
1	1	6%	6%	.14	-1.99	1.8			(- 4.72)		low	low	100%
2	8	50%	56%	-.50*	-.93	1.5	-3.65	1.10	-1.65	-3.68	-3.65	-3.66	78%
3	6	38%	94%	.02	.91	3.0	.31	.66	1.82	.24	.31	.27	69%
4	1	6%	100%	1.86	2.16	1.1	3.34	1.16	(4.45)	3.45	3.34	3.37	100%
									(Mean)		(Modal)	(Median)	

Table 5. Item Measurement Report (arranged by N).

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	N item
63	30	2.1	2.01	1.18	.33	1.1	0	1.2	0	1 gra
76	30	2.5	2.45	-.15	.32	.8	-1	.7	-1	2 dis
93	30	3.1	3.20	-1.90	.33	.6	-1	.6	-1	3 cont
72	30	2.4	2.30	.25	.32	.6	-1	.9	0	4 voc
80	30	2.7	2.61	-.55	.32	1.2	0	1.1	0	5 flu
69	30	2.3	2.20	.56	.32	1.4	1	1.3	1	6 intr
68	30	2.3	2.17	.66	.32	1.3	1	1.2	0	7 soci
75	30	2.5	2.41	-.05	.32	.5	-2	.4	-2	8 all
74.5	30.0	2.5	2.42	.00	.32	1.0	-.4	.9	-.3	Mean (Count : 8)
8.6	.0	.3	.34	.88	.00	.3	1.4	.3	1.2	S.D.

RMSE (Model) .32 Adj S.D. .81 Separation 2.53 Reliability .87
Fixed (all same) chi-square: 56.8 d.f.: 7 significance0: .00
Random (normal) chi-square: 7.0 d.f.: 6 significance: .32

Table 6. Rater Measurement Report (arranged by N).

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	N rater
357	120	3.0	2.99	-1.43	.16	1.0	0	1.0	0	1 A
239	120	2.0	1.94	1.43	.17	.9	0	.9	0	2 B
298.0	120.0	2.5	2.47	.00	.16	.9	-.4	.9	-.4	Mean (Count : 2)
59.0	.0	.5	.53	1.43	.01	.1	.5	.0	.1	S.D.

RMSE (Model) .16 Adj S.D. 1.42 Separation 8.81 Reliability .99
Fixed (all same) chi-square: 157.4 d.f.: 1 significance: .00

Table 7. Comparison of tasks

	N	Mean	S.D.	t	sig.
Task1	5	-.23	1.01		
Task2	2	.61	.05	2.00	<.05

Table 8. Item Measurement Report of Task I

Obsvd	Obsvd	Obsvd	Fair-M		Model	Infit		Outfit		
Score	Count	Average	Average	Measure	S.E.	MnSq	ZStd	MnSq	ZStd	N item
Task 1										
63	30	2.1	2.01	1.18	.33	1.1	0	1.2	0	1 gra
76	30	2.5	2.45	-.15	.32	.8	-1	.7	-1	2 dis
93	30	3.1	3.20	-1.90	.33	.6	-1	.6	-1	3 cont
72	30	2.4	2.30	.25	.32	.6	-1	.9	0	4 voc
80	30	2.7	2.61	-.55	.32	1.2	0	1.1	0	5 flu
76.8	30.0	2.6	2.51	-.23	.32	.9	-.6	.9	-.4	Mean (Count : 5)
9.9	.0	.3	.40	1.01	.01	.2	1.0	.2	.7	S.D.

RMSE (Model) .32 Adj S.D. .96 Separation 2.97 Reliability .90
Fixed (all same) chi-square: 47.3 d.f.: 4 significance: .00
Random (normal) chi-square: 4.0 d.f.: 3 significance: .26

Table 9. Item Measurement Report of Task 2

Obsvd	Obsvd	Obsvd	Fair-M		Model	Infit		Outfit		
Score	Count	Average	Average	Measure	S.E.	MnSq	ZStd	MnSq	ZStd	N item
Task 2										
69	30	2.3	2.20	.56	.32	1.4	1	1.3	1	6 intr
68	30	2.3	2.17	.66	.32	1.3	1	1.2	0	7 soci
68.5	30.0	2.3	2.18	.61	.32	1.4	1.3	1.3	1.0	Mean (Count : 2)
.5	.0	.0	.02	.05	.00	.1	.2	.1	.2	S.D.

RMSE (Model) .32 Adj S.D. .00 Separation .00 Reliability .00
Fixed (all same) chi-square: .1 d.f.: 1 significance: .82

Bibliography

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

Bachman, L.F. and 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., 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.

Brown, J.D. (1996). *Testing in Language Programs.* N.J.: Prentice Hall.

Douglas, D. (2000). *Assessing Languages for Specific Purposes.* Cambridge, UK: Cambridge University Press.

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

Linacre, J.M. (1998). *Facets Rasch Software Users Guide.* 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 and New York: Longman.

O'Loughlin, K. (2001). *The equivalence of direct and semi-direct speaking tests.: Studies in Language Testing 13* Cambridge, UK: 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. (1997a). *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 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 Masters, G.N. (1982). *Rating Scale Analysis: Rasch Measurement.* Chicago, IL: MESA Press.

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