評価者訓練と言語教育のための差異的項目機能の効果的使用法
Effective Use of Differential Item Functioning Information for Rater Training and Language Teaching

中村　優治 NAKAMURA, Yuji

● 慶應義塾大学
Keio University

Keywords 情報，評価者，評価者行動，評価項目，評価者訓練
information, raters, rater performance, evaluation items, rater training

ABSTRACT

This paper will present a study which shows how information regarding different performances by raters on evaluation items can be used to inform rater training and help classroom teachers in their teaching. This study attempts to answer the question: Does a group of raters understand given rating categories and apply them in a similar and consistent way with each other?

本稿では，評価者が評価を行う際に，評価項目が評価基準をどのように共有し，一定の安定した評価を与えているのか差異的項目機能という手法により明らかにすることを試みるものである．この研究の積み重ねにより，評価者の訓練さらには言語クラスの指導に少なからず貢献することが期待される．

1 Theoretical Background and Rationale

Recent research into the performance assessment has employed think-aloud protocols, questionnaires, or interviews to investigate a variety of issues related to rater behavior, providing evidence that different groups of raters behave differently in various ways. There is a need of a systematic investigation of the interpretations raters make of the criteria that are provided for them. The field of language testing is somewhat confused about how far it is reasonable to expect raters to actually agree with each other. Nevertheless, since one of the most commonly stated aims of rater training is to clarify interpretations of the assessment criteria, it seems reasonable to assume that a group of raters should share common interpretations of the features included in a set of specified rating criteria.

Raters in speaking tests are typically examined by inter-rater reliability and intra-rater reliability against the whole test. But this analysis does not necessarily tell us the individual rater's performance on each individual item. The change from the whole test to each evaluation item can provide teachers with the opportunity to use the information for a well-organized rater training and better language teaching.

2 Purpose

This paper will present a study which shows how information regarding different performances by raters on evaluation items can be used to inform rater training and help classroom teachers in their teaching. This study attempts to answer the question: Does a group of raters understand given rating categories and apply them in a similar and consistent way with each other?

3 Method

In this study five raters will be dealt with in their rating behaviors of 12 students' speaking performances using 11 evaluation items (e.g. pronunciation, grammar, eye contact). The students took a speech test and their performances were recorded both on video tapes and audio tapes.

3.1.) Subjects

Twelve Japanese university students

3.2.) Test Material

Speech Test

3.3.) Evaluation items

1. speaker appears sincere to the audience
2. oral fluency
3. pronunciation
4. eye contact
5. facial expression
6. grammar
7. originality
8. content
9. written fluency
10. appropriate evidence
11. holistic evaluation

3.4.) Raters (5)

A,B,C,D,E

3.5.) Rating scale (1- 6 point scale for items 1-10; 1-4 point scale for item 11)

3.6.) Analysis Procedure

In the present study five raters (A, B, C, D, E) judged 12 students' video taped speeches using eleven evaluation items. Although the sample size is relatively small, the results provide us with an example of how eleven evaluation items function differently within a group of raters , and how this can be identified through the Rasch Model.

The acceptable range of the Outfit/ Infit Mean Square in the present study is between 0.6 and 1.4, which is commonly used in the dichotomous data analysis.

4 Results of Basic Measurement Reports

Table 1 provides us with a "birds' eye" view of the three facets (Raters, Students, Items). It tells us that Rater A is the most lenient while Rater B is the harshest among the five raters. It also indicates that Student 8 is the most able one and Student 4 is the least able one. It further shows that Raters were lenient about the eye contact and harsh on the evidence.

In Table 2, The column of Outfit Mean Square (Outfit MNSQ) shows that all of the six rating categories (1 through 6) function appropriately within the acceptable range of between 0.6 and 1.4.

In Table 3, The column of Measure indicates that Rater A is the most lenient and Rater B is the severest among the five of them. The columns of the Infit and Outfit Mean Square provide us with the information that Rater B is misfitting (underfitting in this case) and the other four raters perform appropriately within the acceptable range of between 0.6 and 1.4, though Rater C tends to use the mid part of the rating scale.

In Table 4, the column of Measure shows the

Table 1 Vertical Rulers: Relative Positions among Raters, Students and Items

Measr	+raters	+students	+items
+ 3 +		+	+
		8	
+ 2 +		+ 6	+
		12	
		5 9	
		1	
		3 7	
	A		
+ 1 +		+	+
		11 2	
	D		
		10 4	eye contact pronunciation speaker appears sincere
			oral fluency
			written fluency
			content grammar
* 0 *		*	*
			appropriate evidence facial expression originality
	C E		
	B		
+ -1 +		+	+
+ -2 +		+	+ holistic evaluation

Table 2 Category Function

DATA Category Score	Counts Used	%	Cum. %	QUALITY CONTROL Avge Meas	Exp. Meas	OUTFIT MnSq	STEP CALIBRATIONS Measure	S.E.	EXPECTATION Measure at Category	-0.5	MOST PROBABLE from	.5 Cumul. Probabil. at	Cat PEAK Prob
1	1	0	0	-2.42	-1.47	.6			(-5.41)		low	low	100%
2	27	4	4	-.16	-.48	1.2	-4.29	1.01	-3.05	-4.46	-4.29	-4.36	63%
3	172	26	30	.59	.56	1.0	-1.78	.21	-.67	-1.82	-1.78	-1.80	60%
4	288	44	74	1.25	1.34	1.1	.45	.10	1.40	.42	.45	.43	57%
5	116	18	92	2.01	2.01	1.0	2.60	.10	2.90	2.20	2.60	2.33	37%
6	56	8	100	2.73	2.54	.8	3.03	.16	(4.43)	3.74	3.03	3.40	100%

Table 3 Rater Measurement Report: Examination of raters

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	N raters
611	132	4.6	4.60	1.06	.11	1.27	2.4	1.26	2.2	.64	1 A
576	132	4.4	4.32	.65	.11	.84	-1.4	.86	-1.2	1.12	4 D
498	132	3.8	3.74	-.36	.12	.61	-3.4	.62	-3.3	1.34	3 C
495	132	3.8	3.72	-.40	.12	.74	-2.1	.73	-2.2	1.21	5 E
459	132	3.5	3.46	-.94	.12	1.44	2.9	1.41	2.8	.59	2 B
527.8	132.0	4.0	3.97	.00	.12	.98	-.4	.98	-.3		Mean (Count: 5)
56.5	.0	.4	.42	.74	.01	.32	2.6	.31	2.5		S.D.

RMSE (Model) .12 Adj S.D. .73 Separation 6.24 Separation Reliability .97
Fixed (all same) chi-square: 205.2 d.f.: 4 significance: .00

students in the order of their ability. Student 8 at the top of the table is the most able while Student 4 is the least able.

In Table 5, the column of Measure shows that raters tend to be lenient on "eye contact" and "pronunciation" while they are harsh on "originality" and "evidence." The columns of Infit and Outfit Mean Square indicate that all the evaluation items function appropriately within the acceptable range of between 0.6 and 1.4.

5 Discussion of Differential Item Functioning

Judging from Table 3 above, the five raters with the exception of Rater B perform appropriately within the acceptable range. As a whole, we know that Rater B is misfitting while others are performing as expected. What we still further want to know is how each individual rater is rating on each item, especially how Rater B is behaving on each item compared with the other four. For this purpose, we use the Differential Item Functioning (DIF) theory. We will examine whether each individual rater performs similarly on each of the items. In other words, we will investigate whether each rater performs in the same way for each of the eleven items one by one.

Table 4 Student Measurement Report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Nu	students
246	55	4.5	4.43	2.09	.17	.65	-2.2	.66	-2.0	1.31	8	8
243	55	4.4	4.37	2.00	.17	1.62	2.9	1.68	3.1	.25	6	6
236	55	4.3	4.23	1.79	.17	.87	-.7	.91	-.4	.90	12	12
227	55	4.1	4.07	1.51	.18	1.08	.4	1.13	.7	.87	5	5
227	55	4.1	4.07	1.51	.18	.78	-1.2	.79	-1.1	1.25	9	9
221	55	4.0	3.96	1.33	.18	1.33	1.6	1.36	1.7	.61	1	1
218	55	4.0	3.91	1.23	.18	.96	-.1	.92	-.3	1.16	7	7
217	55	3.9	3.89	1.20	.18	.89	-.5	.91	-.4	1.10	3	3
206	55	3.7	3.70	.84	.18	.82	-.8	.80	-1.0	1.18	11	11
205	55	3.7	3.69	.80	.18	.73	-1.4	.68	-1.7	1.27	2	2
197	55	3.6	3.55	.53	.19	1.15	.7	1.09	.4	.86	10	10
196	55	3.6	3.53	.49	.19	.81	-.9	.81	-.9	1.17	4	4
219.9	55.0	4.0	3.95	1.28	.18	.97	-.2	.98	-.2		Mean (Count: 12)	
16.1	.0	.3	.28	.51	.01	.27	1.4	.28	1.4		S.D.	

RMSE (Model) .18 Adj S.D. .48 Separation 2.68 Separation Reliability .88
Fixed (all same) chi-square: 97.0 d.f.: 11 significance: .00

Table 5 Item Measurement Report: Examination of Items

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Nu	items
257	60	4.3	4.23	.51	.17	.94	-.2	.90	-.5	.92	4	eye contact
256	60	4.3	4.22	.49	.17	.86	-.7	.84	-.8	1.15	1	speaker appears sincer
256	60	4.3	4.22	.49	.17	.68	-2.0	.69	-1.9	1.34	3	pronunciation
250	60	4.2	4.11	.32	.17	.85	-.8	.87	-.7	1.10	2	oral fluency
246	60	4.1	4.05	.20	.17	.90	-.5	.90	-.5	1.22	9	written fluency
244	60	4.1	4.01	.15	.17	.99	.0	.93	-.3	1.10	8	content
243	60	4.1	4.00	.12	.17	1.23	1.2	1.20	1.0	.77	6	grammar
236	60	3.9	3.89	-.09	.17	1.20	1.0	1.21	1.1	.75	5	facial expression
236	60	3.9	3.89	-.09	.17	.73	-1.5	.73	-1.5	1.24	7	originality
235	60	3.9	3.87	-.12	.17	1.13	.7	1.15	.8	.86	10	appropriate evidence
180	60	3.0	2.99	-1.97	.19	1.34	1.7	1.35	1.7	.65	11	holistic evaluation
239.9	60.0	4.0	3.95	.00	.17	.99	-.1	.98	-.1		Mean (Count: 11)	
20.5	.0	.3	.33	.66	.01	.20	1.1	.20	1.1		S.D.	

RMSE (Model) .17 Adj S.D. .64 Separation 3.72 Separation Reliability .93
Fixed (all same) chi-square: 137.3 d.f.: 10 significance: .00

Item 1.

Rater B is lenient and inconsistent in ratings across the student ability range on this item. Along with Rater B, Rater E is also lenient . Rater A is inconsistent, too, but in a strict manner.

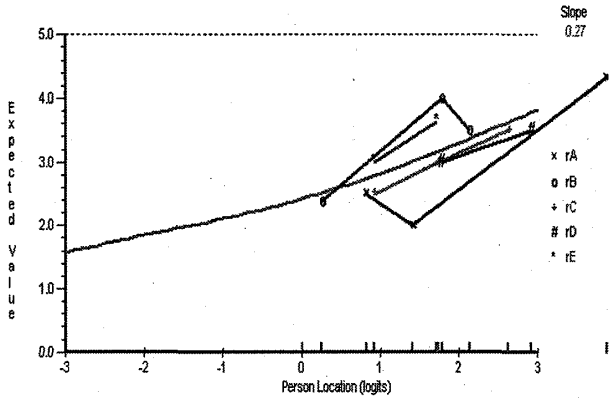

Figure 1 Item: Descriptor for Item 1 [10001]-5Levels for Person Factor: RATER

Item 2.

Rater B is very variable or inconsistent and strict even stricter with low scorers on the test overall than on high scorers. Rater B is not discriminating over low to middle scorers. Raters C and E are also strict on this item.

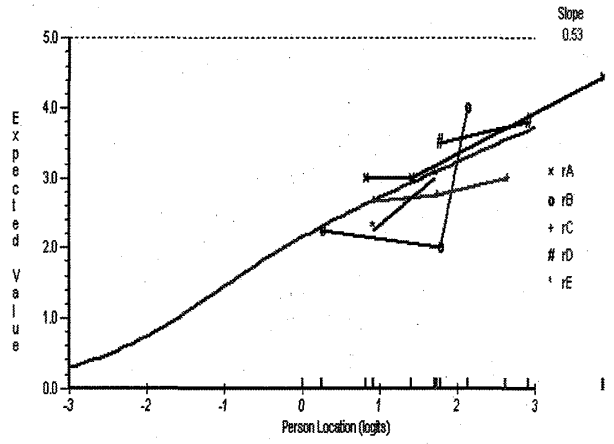

Figure 2 Item: Descriptor for Item 2 [10002]-5Levels for Person Factor: RATER

Item 3.

Rater B is very variable or inconsistent and harsh, even harsher with low level students and less strict with high level students. Rater D is lenient and only very slightly more lenient than others.

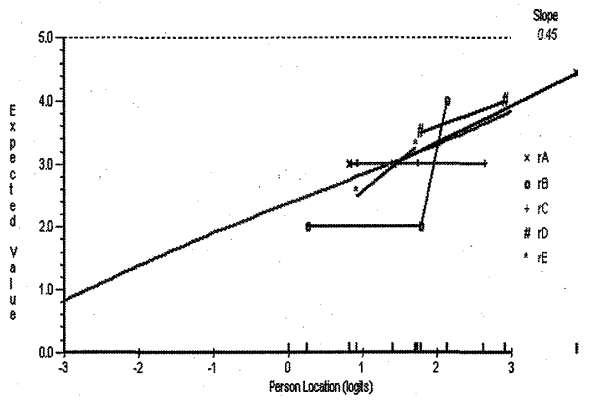

Figure 3 Item: Descriptor for Item 3 [10003]-5Levels for Person Factor: RATER

Item 4.

Rater B is very variable or inconsistent. This rater is stricter with high level students and less strict with low level students. Rater A is strict and consistently stricter across all levels.

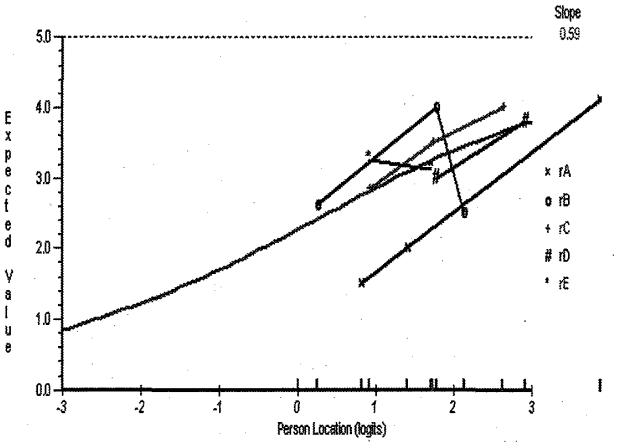

Figure 4 Item: Descriptor for Item 4 [10004]-5Levels for Person Factor: RATER

Item 5.

Rater A is the strictest. Raters B and D are lenient raters. Rater B is rating high level students lower than middle level students on this item.

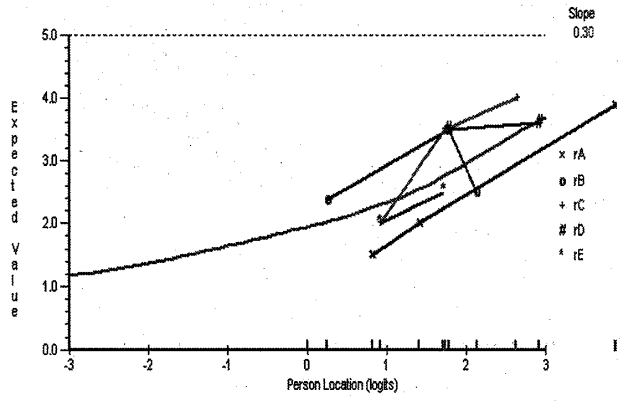

Figure 5 Item: Descriptor for Item 5 [10005]-5Levels for Person Factor: RATER

Item 6.

All raters behave in a greatly similar way on this item. Rater D is lenient marker whereas Rater B is severe.

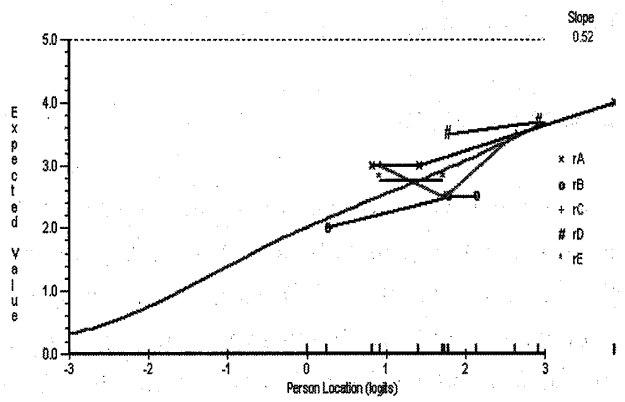

Figure 6 Item: Descriptor for Item 6 [10006]-5Levels for Person Factor: RATER

Item 7.

On the whole, all the raters are quite similar to one another on this item. Rater A is the most severe and severest on low to mid range level students. Rater B is variable or inconsistent.

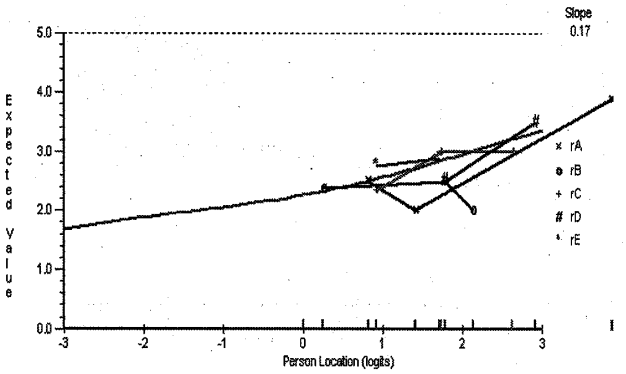

Figure 7 Item: Descriptor for Item 7 [10007]-5Levels for Person Factor: RATER

Item 8.

Rater A is the most lenient. Rater D is a little strict, but consistent across all students.

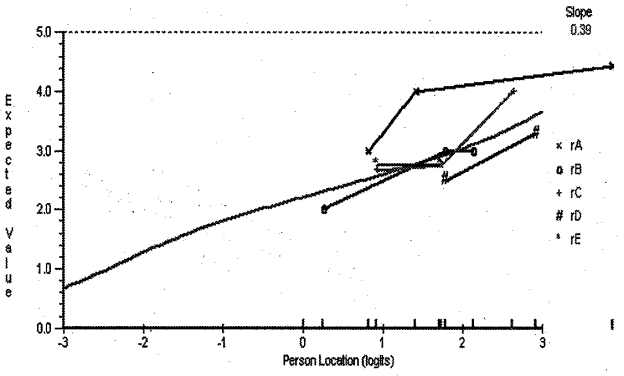

Figure 8 Item: Descriptor for Item 8 [10008]-5Levels for Person Factor: RATER

Item 9.

Rater A is the most lenient. Rater B is variable or inconsistent and Rater B overdiscriminates between middle and high level students.

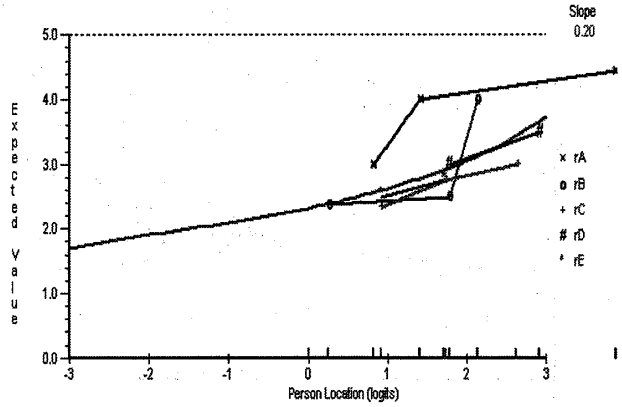

Figure 9 Item: Descriptor for Item 9 [10009]-5Levels for Person Factor: RATER

Item 10.

All the raters are quite similar in their ratings on this item. Rater B is the most lenient. Rater A is a bit lenient on low level students

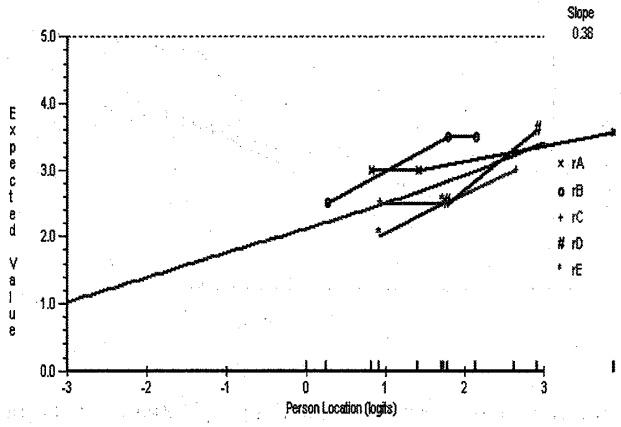

Figure 10 Item: Descriptor for Item 10 [10010]-5Levels for Person Factor: RATER

Item 11.

Raters B, C and E are lenient.. Rater D is the strictest. Rater A is slightly inconsistent.

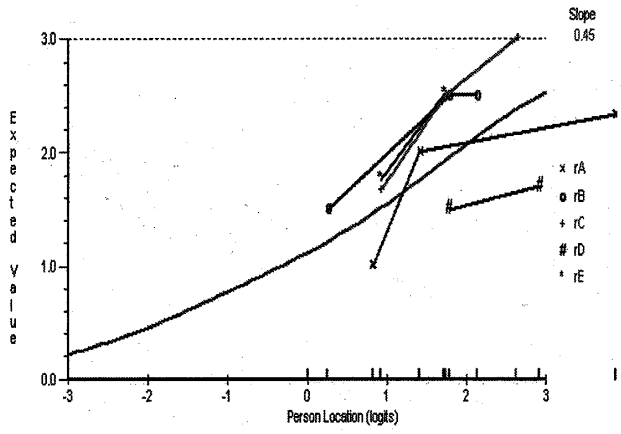

Figure 11 Item: Descriptor for Item 11 [10011]-5Levels for Person Factsor: RATER

Based on the results of the DIF analysis , we can draw a summary as follows: Overall, we say that five raters are similar on most items, especially items 6,7 and 10. However, as was pointed out in the misfitting rater section, Rater B is inconsistent in several items, for which there may be a reason. Rater A sometimes behave inconsistently, but still is within the acceptable range, as was described in Infit and Outfit Mean Square section.

6 Conclusion

The conclusions we could draw from the present data analysis are as follows: We should answer the research question, "Does a group of raters understand given rating categories and apply them in a similar and consistent way with each other? "

1) On the whole, the present group of raters functions appropriately to rate the students' speaking ability. 2) The results, although the sample might be too small to make a broad generalization, show that each individual rater's characteristic investigation through individual items can be more

useful than the whole statistical analysis of inter-rater reliability for the rater training when we have high stakes tests.

Traditionally, differences in performance among raters have been examined by way of the inter-rater reliability. However, through this DIF theory it can be pointed out that each individual rater could perform differently on an item. The emphasis can be shifted from the group of raters to each individual rater, and even from the tests to the items themselves.

References

Andrich, D. & Styles, I. (2004). *Report on the Psychometric Analysis of the Early Development Instrument (EDI) Using the Rasch Model.* Centre for Learning , Change and Development. School of Education, Murdoch University.

Andrich, D., Sherican, B. & Luo, G. (2004). *RUMM 2020: a Windows Program for the Rasch Unidimensional Measurement Model.* Perth, Western Australia: RUMM Laboratory