ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 6, No. 4, pp. 866-872, July 2015
DOI: http://dx.doi.org/10.17507/jltr.0604.21

Validation of an Oral English Test Based on Many-faceted Rasch Model*

Shujing Wu
Binzhou University, Binzhou City, China

Tongpei Dou
Binzhou University, Binzhou City, China

***Abstract*—This Study investigates the validity of an English Oral English Test from three aspects: raters, examinees, and task difficulty based on the theory of Many-faceted Rasch Model by using FACETS. The results show that there exist significant differences in the examinees' oral ability and that raters' rating has good internal self-consistency, but there is significant difference in their severity and that tasks are significantly different in difficulty and that the differentiation is good enough to distinguish the examinees' ability. In general, the good validity of the Oral English Test is provided, but the process of the examinees' performance needs further study.**

***Index Terms*—oral English test, validity, Many-faceted Rasch Model**

I. Introduction

Validity is used to interpret the appropriateness of giving tests and argue the rationality and sufficiency of the test scores (Messick, 1989; APA, 1999; Zou, 2005).The research of validity is the process of verifying the various inferences and behavioral decision-makings for the test scores, based on the theoretical and empirical evidence. According to the "*Standards for Educational and Psychological Testing*" (APA, 1999), the main sources of evidence for the validity are from five aspects: test content; reaction process; the internal structure of the test; the relationship between the test scores and other external variables; consequences of the tests (Zou, 2005). The reaction process constitutes two aspects which are the candidates' psychological reaction process when their taking the tests and the raters' psychological reaction when their scoring subjective items (Jiang & Wen, 2010). Rating validity is a primary evidence to examine the efficiency of a performance assessment (Weir, 2005; Bachman, 2004). Hoyt & Kerns (1999) argue that more than one third of the test score differences are caused by the rater effects and interaction between the examinees and raters. Therefore the study on the rater effects is an important prerequisite for the interpretation for the validity of performance assessment. Most of researches on the validity of oral tests are focused on test content design, rating criteria, and organizational forms. However, few are related to the reaction process of the examinees and raters in China (Jiang & Wen, 2010).

In recent years, many researches on the language testing have been carried out based on the Many-facet rasch measurement model (Eckes, 2005; Elder et al., 2007; Jiang & Wen, 2010). Many-facet rasch measurement model is one of the measurement models of Item Response Theory and it can be realized by FACETS statistical analysis software. Many-facet rasch measurement model is prior to other measurements because it can provide test-free, scale-free and sample-free calibration of items, and the judgment for the accuracy of rating criteria and determination whether there are significant differences between the internal components of the facets and whether there is interaction between different facets. However, compared with logical analysis, correlation analysis, questionnaire and interview, this model is seldom used in the study on the validity of language testing by researchers in China (Jin & Guo, 2002; Huang, 2006; Wang, 2007). Until now, only few researchers have used many-facet rasch measurement model to make a study on the validity of some item types, such as discourse cloze test and translation test (Liu, 2005; Jiang & Wen, 2010).

II. Research Design

This study aims to use many-facet rasch measurement model to explore the validity of an oral test by exploring the psychological reactions of the examinees and raters. If the data values of misfit validity are within the acceptable limit, it indicates that the test has a high fit validity (Linacre, 2008; Jiang & Wen, 2010).

A. *Research Questions*

The purpose of the study is to examine the validity of an oral test. The specific research questions are as follows:

(1) Are the examinees' response behaviors self-consistent in the oral test?

* Supported by Shandong Education Sciences Planning Fund Program (Program No. ZK1123332B020).

(2) Are task difficulties reasonable enough to distinguish the examinees' oral performances in the oral test?
(3) Are the raters' internal rating behaviors consistent in the oral test?

B. *The Participants*

The English oral test is part of the university entrance examination which intends to provide references for the universities to select the talents by examining whether the examinees' oral performances have met the requirements of *The New English Curriculum Standards for Senior High School* (2007) and whether they can fulfill the tasks by applying their acquired knowledge and skills. Three item types were designed in the oral test which was "Reading aloud", "Answering questions" and "Free conversation". The face-to-face oral test was used. The trained raters were required to rate the examinees' oral performances according to the rating criteria for the tasks which consist of five scales (content, intonation and pronunciation, vocabulary, grammar, and communicative strategies). The analytical approach was used. Each examinee was rated by two raters, and the two scores were averaged. In cases of extreme score differences, a third rater was required and the two scores were close to each other used as the final score. Three hundred examinees from different senior high schools and twelve raters from the universities in Shandong province were chosen as subjects. The raters had more than three-year teaching and rating experience, with 5 males and 7 females.

C. *Data Collection*

All the examinees took the oral test in June, 2014 and were required to complete the three tasks within 15 minutes. The two raters for each group independently rated the performance of each examinee according to the rating criteria. The maximum mark is 100.

III. Results and Discussion

The oral abilities of the examinees, raters' scores and tasks were defined as three facets in the study and many-facet rasch measurement model analysis for the oral test was achieved by using the software package FACETS (Linacre, 2008) in this study.

A. *The True Measurement Value of Each Facet*

Table 1 is a descriptive summary that shows the true measured values of each facet without the effect of other facets. The scale along the left side of Table 1 represents the logit scale, ranging from +4 to -3. The "Measure" in the first column of the table represents the individual scale values based on the same measurement unit "logit", which facilities the comparison and analysis of each facet. The second column of the table represents the examinees' oral performance values with the highest oral quality performer at the top and the lowest oral quality performer at the bottom, and each asterisk represents four examinees and each dot represents less than four examinees. As can be seen from the table, the values of examinees' oral abilities are ranging from -2 to +3. The third volume of the table refers to raters' severity, which is ordered in accordance with the level of severity, with the most severe raters at the top and the lenient raters at the bottom. According to Table 1, the severity value range of the raters is between -1 and +1, and the distribution is relatively concentrated, indicating that the test scores given by the raters are more consistent. The fourth column is task difficulty, according to a top-down arrangement of task difficulty. Table 1 shows the values of task difficulties are in the range of -1 and +1, with moderate difficulty, and Task 3 "Free conversation" is more difficult than Task 1 "Reading aloud" and Task 2 "Answering questions". The fifth column is the examinees' estimated score values, and the examinees with the logit zero should get roughly 84.5 points, and the highest score that the examinees obtain is 98 and the lowest score 62.

TABLE 1.
EXAMINEES, RATERS, AND TASKS SUMMARY REPORTS

Measr	examinees	-raters	-difficulty	Scale
4		+		(98)
		|		
	.	|		---
		|		96
	.	|		
		|		---
	.	|		95
3	*	+		
	***.	|		---
	.	|		94
	*.	|		93
	*	|		---
	.	|		92
	.	|		91
2	.	+		90
	*.	|		---
	*.	|		
	*.	|		89
	***	|		---
	.	|		
		|		88
1	*.	+		
	.	|		---
	.	|		87
	.	|	Task 3	---
	****.	|		
	******.	|		86
	*********.	R1 R11 R12 R2 R5 R6		85
* 0	* ********.	* R10 R9	*	* --- *
	******.	R3 R4 R8		84
	*****.	R7	Task 1 Task 2	83
	**.	|		---
	*.	|		82
	.	|		81
	**.	|		---
-1	.	+		80
	.	|		79
	.	|		78
	.	|		77
		|		75
		|		74
	.	|		71
-2	*.	+		70
		|		69
	.	|		68
		|		66
		|		---
		|		
		|		65
-3	+	+		(62)
Measr	* = 4	-raters	-difficulty	Scale

B. *Examinees' Oral Performance Analysis*

Table 2 shows the overall measure of the examinees' oral performances. In other words, it presents the examinees' oral ability. The average of the actual examinees' score given is 85.45, and the average of the estimated score calculated by Rasch model is 85.54, with a difference of 0.09, which indicates the average score actually given by the raters is roughly the same as estimated. Separation represents the differences of examinees' abilities, and larger values indicate greater differences between the examinees' abilities. If the value of separation is more than 2, it means that there exists a significant difference between the individual examinees. The value of separation in Table 2 is 5.60, showing that there are significant differences between the examinees' abilities in this text. "Reliability" here refers to the reliability of separation index, instead of inter-rater reliability. Cronbach ranges from 0 to 1, and the larger value means that the greater difference between the candidates' abilities. The reliability of separation index in Table 2 is 0.97, indicating that there is a great difference between the examinees' abilities. The Chi-square test can be used to examine the significant degree of the differences so that the judgment can be made statistically on the differences between the examinees' abilities. In table 2, the Chi-square value is significant at p=.00, showing there are significant differences between the examinees’ abilities, which is the requirement for large-scale examinations.

TABLE 2.
EXAMINEES’ MEASUREMENT REPORT

Mean	Observed average	85.45
	Fair（Mean）average	85.54
Separation	Separation	5.60
	Reliability	.97
Chi-square	significance(probability)	.00

TABLE 3.
THE CASES OF THE EXAMINEES’ PERFORMANCE REPORT

Examinee	Observed average	Fair (M) average	Measure	Infit MnSq	ZStd	Outfit MnSq	ZStd
21	96.00	96.30	3.68	.47	-.7	.44	-1.0
151	96.00	95.68	3.37	.43	-.8	.41	-1.0
270	84.50	85.10	.14	.38	-1.2	.38	-1.2
1	83.17	83.66	-.17	.19	-2.0	.19	-2.0
245	71.50	70.75	-1.94	.23	-1.7	.32	-1.3
79	69.83	67.49	-2.25	.40	-1.0	.41	-.8

Table 3 shows the data for the cases of the examinees’ performances. The columns of the table from the first to the eighth are as follows: the numbers of the examinees, the average scores given by the raters, the true values calculated by the rasch model, the true values of the examinees’ abilities, the weighted mean square fit statistics, standard fit data with the normal distribution, conventional (unweight) mean square fit statistics, standard fit data with the normal distribution (unweight). The two fit statistics--infit and outfit in the fifth column and seventh column show the consistency of the examinees’ individual behavior (Linacre, 2008). High infit statistics are a little more problematic compared with high outfit statistics which are more sensitive to extreme scores. Linacre recommends that its critical range should be between 0.5 and 1.5. The statistics in the sixth column and the eighth column are supplement to the fit statistics in the fifth column and seventh column, whose absolute value is less than 2 or 3, indicating the examinee individual behaviors are fit for the Rasch model (Linacre, 2008). In Table 3, the fit statistics show that the examinees’ performances fit the model and infit and outfit values are within the acceptable range (0.5-1.5) and that the separation index of the measure (5.60, in Table 2) exceeds the minimum limit of the acceptable score 2.0 to separate the examinees’ oral abilities.

TABLE 4.
EXAMINEES ABILITY FIT STATISTICS

Fit range	No. of examinees	Infit MnSq	No. of examinees	Outfit MnSq
Fit<0.5	19	6.33%	19	6.33%
	11	3.67%	9	3.00%
0.5≤Fit≤1.5	232	77.33%	234	78.00%
Fit>1.5	23	7.67%	23	7.67%
	15	5.00%	15	5.00%

The examinees’ abilities fit statistics are shown in Table 4, with an average score 84.5 of the examinees as a critical line. The examinees whose scores were higher than 84.5 points were regarded as high ability examinees and lower than that score were considered as low ability examinees. Based on Table 4, 7.67% higher ability examinees’ weighted mean square fit values were greater than 1.5. The possible reason might be that they were too nervous or showed contempt for the test, resulting in an unexpected loss of points. 5.00% of the lower ability examinees’ weighted mean square fit values were greater than 1.5. They might be in good mental state in the test or the topics in the test were quite fit for them, leading to their better performances and gained unexpected scores. Another 10% of the examinees’ weighted mean square fit values were less than 0.5. It was probable that they took indifferent attitudes towards the test and slipped off during the test. In general, 22.67% of the examinees’ actual performances were not consistent with the estimated ability of the examinees that was beyond the acceptance limit 2.00%, which indicates that the occurrence of the self-inconsistencies in the examinees’ behaviors.

The misfit statistics about the examinees, as well as the corresponding task and raters are shown in Table 5. Data indicates that the internal inconsistencies in the examinees’ response behaviors were closely related to Task 1 “Reading aloud” and Task 3 “Free Conversation”. What happened to the examinees when they were doing the tasks and what psychological factors caused their different behaviors, which need deeply qualitative analysis to find out the reasons for their better performances and worse performances through interviews, observations, questionnaires, thinking aloud and other qualitative data. If necessary, the deviation analysis was to be done that was beyond the scope of the validity study, which was not touched in the paper.

TABLE 5.
MISFIT STATISTICS OF THE EXAMINEES

Cat	Score	Exp.	Resd	StRes	Num	exam	Nu	rat	N	criter
70	67	77.3	-10.3	-5.0	237	S237	10	R10	1	Task 1
70	67	77.2	-10.2	-4.9	7	S7	2	R2	1	Task 1
70	67	77.2	-10.2	-4.9	107	S107	6	R6	1	Task 1
70	67	77.2	-10.2	-4.9	257	S257	12	R12	1	Task 1
94	90	84.1	5.9	4.4	82	S82	4	R4	3	Task 3
94	90	84.1	5.9	4.4	177	S177	8	R8	3	Task 3
94	90	84.1	5.9	4.3	47	S47	2	R2	3	Task 3
94	90	84.1	5.9	4.3	147	S147	6	R6	3	Task 3
94	90	84.1	5.9	4.3	212	S212	10	R10	3	Task 3
94	90	84.1	5.9	4.3	297	S297	12	R12	3	Task 3
81	77	83.2	-6.2	-4.1	206	S206	10	R10	1	Task 1
81	77	83.2	-6.2	-4.0	41	S41	2	R2	1	Task 1
81	77	83.2	-6.2	-4.0	76	S76	4	R4	1	Task 1
81	77	83.2	-6.2	-4.0	141	S141	6	R6	1	Task 1
81	77	83.2	-6.2	-4.0	291	S291	12	R12	1	Task 1
81	77	83.1	-6.1	-3.9	171	S171	8	R8	1	Task 1
83	79	84.1	-5.1	-3.8	47	S47	1	R1	3	Task 3
83	79	84.2	-5.2	-3.8	82	S82	3	R3	3	Task 3
83	79	84.1	-5.1	-3.8	147	S147	5	R5	3	Task 3
83	79	84.2	-5.2	-3.8	177	S177	7	R7	3	Task 3
83	79	84.1	-5.1	-3.8	212	S212	9	R9	3	Task 3
83	79	84.1	-5.1	-3.8	297	S297	11	R11	3	Task 3
86	82	71.6	10.4	3.2	7	S7	1	R1	3	Task 3
86	82	71.6	10.4	3.2	107	S107	5	R5	3	Task 3
86	82	71.5	10.5	3.2	237	S237	9	R9	3	Task 3
86	82	71.6	10.4	3.2	257	S257	11	R11	3	Task 3
88	84	77.7	6.3	3.0	6	S6	2	R2	3	Task 3
88	84	77.7	6.3	3.0	106	S106	6	R6	3	Task 3
88	84	77.7	6.3	3.0	236	S236	10	R10	3	Task 3
88	84	77.7	6.3	3.0	256	S256	12	R12	3	Task 3

C. *The Raters' Severity and Their Internal Consistency*

The performances of the raters were analyzed mainly from two aspects: the raters' severity and the raters' internal consistency in this study.

TABLE 6.
RATERS' MEASUREMENT REPORT

Obsvd	Fair(M)	Model		Infit		Outfit		Estim.	Nu raters
Average	Average	Measure	S.E.	MnSq	ZStd	MnSq	ZStd	Discrm	
84.66	85.89	.13	.04	1.29	2.2	1.36	2.7	.25	2 R2
84.66	85.89	.13	.04	1.29	2.2	1.36	2.7	.25	6 R6
84.66	85.89	.13	.04	1.29	2.2	1.36	2.7	.25	12 R12
84.73	85.96	.11	.04	.95	-.3	.89	-.9	.40	1 R1
84.73	85.96	.11	.04	.95	-.3	.89	-.9	.40	5 R5
84.73	85.96	.11	.04	.95	-.3	.89	-.9	.40	11 R11
85.19	86.31	.01	.04	.95	-.3	.89	-.9	.76	9 R9
85.21	86.32	.01	.04	1.28	2.0	1.36	2.6	.55	10 R10
86.87	86.81	-.15	.04	.89	-.9	1.01	.1	.71	4 R4
86.33	86.83	-.16	.04	.91	-.7	1.01	.1	.43	8 R8
87.03	86.93	-.20	.04	.54	-4.4	.60	-3.7	.76	3 R3
86.57	87.00	-.22	.04	.54	-4.4	.60	-3.7	.41	7 R7
85.45	86.31	.00	.04	.99	-.3	1.02	.0	.90	Mean (Count: 12)
.92	.44	.14	.00	.26	2.2	.27	2.3	.01	S.D. (Population)
.96	.45	.14	.00	.27	2.3	.29	2.4	.01	S.D. (Sample)

Model, Populn: RMSE .04 Adj (True) S.D. .13 Separation 3.21 Strata 4.61 Reliability .91
Model, Sample: RMSE .04 Adj (True) S.D. .14 Separation 3.36 Strata 4.82 Reliability .92
Model, Fixed (all same) chi-square: 131.7 d.f.: 11 significance (probability): .00
Model, Random (normal) chi-square: 10.2 d.f.: 10 significance (probability): .43

At the raters' facet, "Measure" and "Infit" are used to interpret raters' individually internal reliability. That is to say, the severer the raters are and the higher the corresponding "Measure" values are. The more lenient the raters are and the lower the corresponding "Measure" values are (Shi & Han, 2009). As is shown in Table 6, Rater 2, Rater 6 and Rater 12 were the severest and Rater 7 was the most lenient of all the raters. The separation between the raters in Table 6 was 3.21 and reliability of separation index was 0.91 and the chi-square value was significant at p=.00, indicating that there was a significant difference in raters' severity.

According to the values of Infit MnSq in the table, the infit values of Rater 2, Rater 6, Rater 12, and Rater 10 were greater than 1, indicating the rating variations by the four raters were larger than estimated by the Rasch model, which was not fit enough for the model while the other eight raters were quite different whose fit values were less than 1, which indicated their rating behaviors were overfit for the model, whose rating variations were less than estimated by the Rasch model. However, “Infit” values were within the acceptable ranges (0.5-1.5), and central tendency or polarization phenomenon did not occur in the 12 raters’ ratings, which showed that raters did distinguish the examinees’ abilities reasonably. Therefore it was safe to say that there was great internal consistency in their ratings.

D. *Task Difficulty*

In order to learn about the differentiation of the test, the difficulties of the tasks and the rating difficulties were analyzed. (See Table 7)

TABLE 7.
TASK DIFFICULTY MEASUREMENT REPORT

Obsvd	Fair(M)	Model		Infit		Outfit		Estim.	N tasks
Average	Average	Measure	S.E.	MnSq	ZStd	MnSq	ZStd	Discrm	
82.93	83.94	.56	.02	1.21	3.2	1.26	3.8	.82	3 Task 3
86.69	87.14	-.28	.02	.63	-6.8	.64	-6.7	.45	2 Task 2
86.72	87.16	-.29	.02	1.10	1.5	1.15	2.3	.10	1 Task 1
85.45	86.08	.00	.02	.98	-.7	1.02	-.2	.89	Mean (Count: 3)
1.78	1.52	.40	.00	.25	4.4	.27	4.7	.02	S.D. (Population)
2.18	1.86	.49	.00	.31	5.4	.33	5.7	.02	S.D. (Sample)

Model, Populn: RMSE .02 Adj (True) S.D. .40 Separation 19.70 Strata 26.60 Reliability 1.00
Model, Sample: RMSE .02 Adj (True) S.D. .49 Separation 24.14 Strata 32.52 Reliability 1.00
Model, Fixed (all same) chi-square: 1234.8 d.f.: 2 significance (probability): .00
Model, Random (normal) chi-square: 2.0 d.f.: 1 significance (probability): .16

Based on Table 7, separation of the tasks was 19.70 and the reliability was 1.00 and the chi-square value are significant at p=.00, indicating that there were significant differences in the task difficulty, which was the basic features of tests. The values of the tasks (logit) in the column “Measure” showed the rating difficulties of the tasks. Task 3 “Free conversation” was the most difficult and the raters rated more severely in this task and it was difficult for the examinees to get more scores than estimated. And the second one was Task 2 “Answering questions” in the aspect of rating difficulty. And it was easier for the examinees to get marks in Task 1 “Reading aloud” and the raters were lenient in their ratings. However, the values of “Infit MnSq” were within the acceptable range (0.5 to 1.5), indicating there was good discrimination between the three tasks.

IV. CONCLUSION

This study analyzed the validity of an oral test from the three facets which were examinees, raters and task difficulty through many-facet Rasch Measurement Model. The results showed that there existed significant differences in the examinees’ oral ability and that raters’ rating had good internal self-consistency, but there was significant difference in their severity and that tasks were significantly different in difficulty and that the differentiation was good enough to distinguish the examinees’ ability. In general, the good validity of the Oral English Test was provided, but the process of the examinees’ performance needs further study.

The study has a positive role in the research on oral English tests, which can make up for the lack of the study on the validity of oral tests through many-facet measurement model, which can be used to examine the effects of various factors on the tests and can be widely applied in language testing. It is hoped that more scholars devote to the related research.

REFERENCES

[1] American Psychological Association, and National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

[2] Bachman, L. F. (2004). Statistics Analyses for Language Assessment. Cambridge: Cambridge University Press.

[3] Eckes, T. (2005). Examining rater effects in TestDaf writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly* 2. 3, 197-221.

[4] Elder, C., Barkhuizen, G., Knoch, U. & Von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing* 24. 1, 37-64.

[5] Hoyt, W. T., & Kerns, M. D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods* 4, 403–424.

[6] Huang, Yonghong. (2006). Validity and Reliability of TEM4-Oral. *Foreign Languages Research* 3, 36-38.

[7] Jiang, Jin-lin., & Wen, Qiu-fang. (2010). Validation of a Translation Exam Based on Rasch Measurement Model. *Computer-assisted Foreign Language Education* 131, 14-18.

[8] Jin, Yan,. & Guo, Jieke. (2002). The Validity of CET Semi-Direct Oral Proficiency Test. *Foreign Language World* 5, 73-79.

[9] Linarcre, J. M. (2008). A user’s guide to FACETS: Rasch-model Computer Program. Chicago: MESA Press.

[10] Liu, Jianda. (2002). Testing Approaches for Discourse Cloze Test Based on multidimensional Rasch model analysis. *Modern Foreign Languages* 28. 2, 51-63.

[11] Messick, S. (1989). Validity. In R. L. Linn (Ed.). *Educational Measurement* (3rd ed.). New York: Macmillan.

[12] Shi, Baoliang. & Han, Baocheng. (2009). The Application of Many-facet Rasch Analysis Software Facets in English Testing.

English Language Education in China 2, 1-10.

[13] Wang, Haizhen. (2007). Validation of TEM4-Oral: Evidence from Raters Assessment Process. *Journal of PLA University of Foreign Languages* 30. 4, 49-53.

[14] Weir, C. J. (2005). Language testing and validation: An evidence-based approach. Houndmills, UK: Palgrave Macmillan.

[15] Wen, Qiufang. & Zhao, Xuexi. (1998). The theory and practice for the assessment of Tape-Mediated TEM4-Oral. *Journal of PLA University of Foreign Languages* 2, 52-55.

[16] Zou, Shen. (2005). Language Testing. Shanghai: Shanghai Foreign Language Education Press.

Shujing Wu is an associate professor at Binzhou University. She is interested in TEFL, language testing, and teacher education.

Tongpei Dou is an associate professor at Binzhou University. He is interested in applied linguistics.