On the Construct Validity of an Analytic Rating Scale for Speaking Assessment

Chunguang Tian[1,2,*]

[1]Foreign Languages Department, Binzhou University, Binzhou, P.R. China

[2]English Education Department, School of Education, Chonbuk National University, Jeonju si, Korea

*Correspondence: English Education Department, 567 Baekje-daero, deokjin-gu, Jeonju-si, Jeoolabuk-do 561756, Republic of Korea. Tel: 10-2918-0086. E-mail: bztcg@bzu.edu.cn

Received: July 8, 2016 Accepted: July 21, 2016 Online Published: September 19, 2016

doi:10.5430/ijelt.v4n1p1 URL: http://dx.doi.org/10.5430/ijelt.v4n1p1

Abstract

The analytic rating scale is often used in the assessment of learners' speaking ability. Compared with holistic rating scale, the analytic rating scale can provide much more information about the test takers. But the studies in this field are not so fruitful. This paper aims to study the construct validity of an analytic rating scale for speaking assessment. The Multi-facet Rasch Modeling method and the correlations analysis are combined to investigate the construct validity. The MFRM analysis shows that there is a good reliability between raters in terms of severity and consistency; the analytic rating scale can reflect the students' speaking ability. But the correlation analysis indicates that there is no good discriminant validity for the four rating criteria, but excellent convergent validity. This study gives some implication in the design of rating criteria and the rater training as well.

Keywords: Construct validity, analytic rating scale, speaking assessment, multi-facet Rasch analysis

1. Introduction

Analytic rating scales are usually adopted to judge students' language ability in a single modality (for instance, speaking) in the field of second language assessment because they incorporate large amount of information about students' language competence, hence preferable over holistic rating scales (Brown & Bailey, 1984; Pollitt & Hutchinson, 1987; Kondo-Brown, 2002; Bachman, Lynch & Mason, 1995). For that reason, it is necessary to attract the raters to focus on the specific rating criteria to improve the rating accuracy (Brown & Bailey, 1984; Luoma, 2004) and consistency within the framework of multidimentional definition of language competence (Bachman, Lynch & Mason, 1995).

The exploration of analytic scales is often based on certain theories of language abilities and the intended testing purposes (Luoma, 2004). So it is necessary to study its validity after exploration from different perspectives. Construct validity is the most important one in testing validity, referring to the extent to which performance on tests is consistent with predictions that we make on the basis of a theory of abilities or constructs (Bachman, 1990:255). For the analytic rating scales in speaking test, there are three empirical evidences to support the test design: 1) the raters' consistency in interpreting and using the analytic rating scales; 2) the analytic scales can distinguish the examinees' language ability; 3) the different analytical rating scales are supposed to be closely related each other on the one hand and to be distinct enough to prove that the analytical rating scales could reflect different aspects of an individual's language competence on the other hand. (Sawaki, 2007)

When it comes to these crucial issues, several previous studies have touched upon the construct validity of analytic rating scales, though only a few have attempted to investigate convergent and discriminant validity, and the relationships between analytic ratings and overall score. The current study will look into such issues by combing Multi-facet Rash Model (MFRM) and Correlation analysis, hoping to bridge the gap. This paper tries to answer the following questions:

1) How are the raters' severity and consistency in rating? What about the rating validity?

2) Is it appropriate to use 9-scales for the rating criteria? Can the rating scales distinguish the examinees' language ability effectively?

3) How are the convergent validity and discriminant validity of the rating scale?

2. Literature Review

The issue of validity has received great attention in L2 assessment (Luoma, 2004). The traditional studies focused on the inter-rater and intra-rater consistency, which can not reflect the real rating accuracy even if there were high correlation, because it couldn't reflect the raters' correct use of rating scales (Eckes, 2011). The rating process involves various factors besides the aspect of the rater. In L2 performance assessment, particularly in speaking and writing tests, the rating validity is of great significance, which is determined by many factors including raters, rating criteria, the rating process etc (Weir, 2005). The traditional studies only took into account of the issue of raters, but not such other factors as task difficulty, rating criteria, etc (LIU, 2005). In recent years, the use of Multi-facet Rasch Model has been widely used in language performance assessment because of its providing a relatively reliable value by taking into consideration various factors in language testing. Eckes (2005), HE Lianzhen and ZHANG Jie (2008) studied the rating validity of speaking tests in GFL and CET. Their studies showed different degrees in raters' inconsistency despite holistic consistency.

In spite of the fact that there are many research on raters' consistency, the previous studies gave little evidence for the interrelationships, or convergent/ discriminant validity among analytic rating scales. This issue was investigated mainly within two theoretical framework: one is the factor analysis, combing with language ability measures of other modalities to prove the effectiveness of ratings (e.g., Bachman, Davison, etc, 1995; Carroll, 1983; Shin, 2005); the other is the multivariate G theory to language assessments which reported the analytic rating scales were highly correlated with each other (e.g. Lee, 2005; Lee & Kantor, 2005; Sawaki, 2003, 2005; Xi, 2003).

There are little studies on the relationship of analytic rating scales to the overall score. McNamara (1990) and Elder (1993) did such studies on this topic, who obtained a separate overall rating and analytic rating scores for different aspects of language competence. In McNamara's study, an unexpected interdependence between Overall Effectiveness and Resources of Grammar and Expression was found, with the speculation that this could be explained by the role of grammar in the allocation of scores by raters. The followed stepwise regression also supported the above finding. Elder (1993) did the similar study, finding content specialist and ESL specialist weighed different analytic scales differently. Weigle's (1998) and Wang and Stanley's (1970) studies showed how to obtain a composite score by different weighing of analytical rating scales.

The current study, the methods of MFRM and SPSS correlation analysis are combined to investigate the construct validity of analytic rating scales in speaking test. On the one hand, the MFRM can provide information about every facet in the process of testing including examinees' language ability, raters' severity and task difficulty, and the individual information of examinees and raters; on the other hand, the SPSS correlation analysis was employed to examine the correlation among the four rating scales.

3. Research Design

3.1 Participants

Twenty eight junior undergraduates of the same class from English department of Chonbuk National University were randomly selected as the participants, who attended the spoken English test. The students were not supposed to open the question paper in advance to make any preparation.

3.2 Raters

The raters were 2 Ph. D graduates majoring in English language education with rich language teaching experience. The raters training process was carried out by familiarizing the raters with the speaking test form and rating scales. To guarantee the best training result (rater consistency), some speaking samples were provided for training exercises. Each rater then rated 28 recorded audio files at home separately.

3.3 Data Collection Analysis

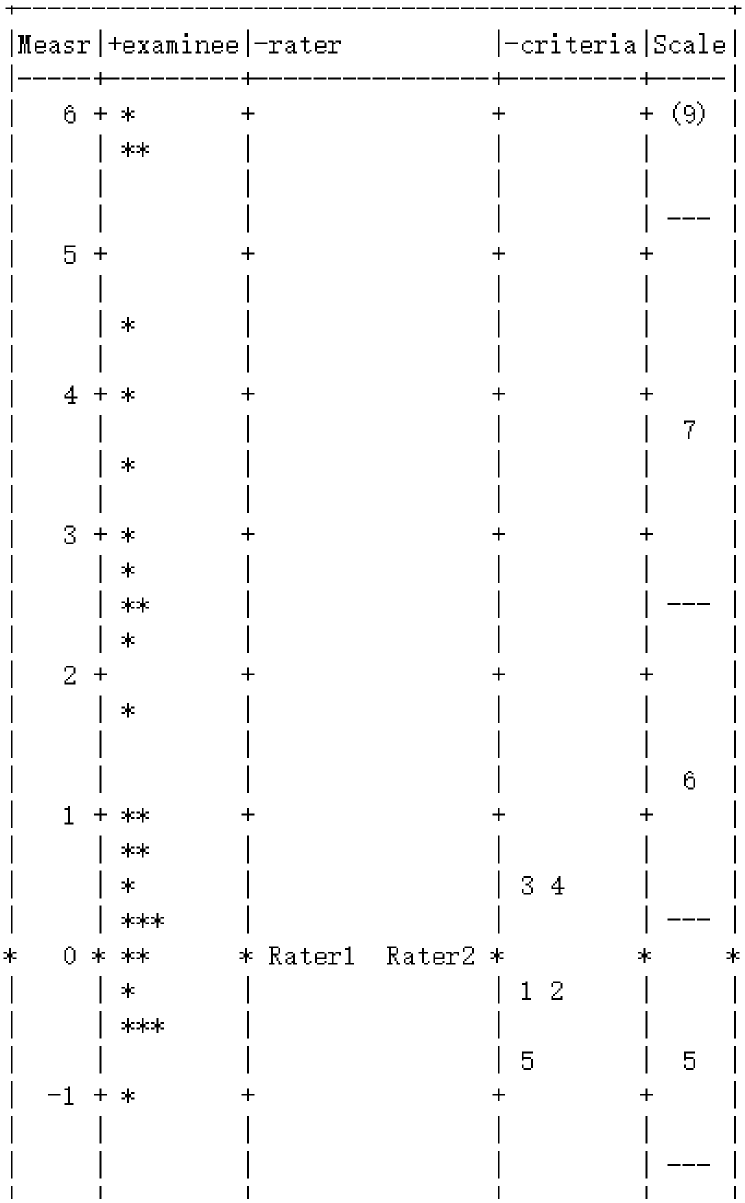
The test took the form of instantaneous monologue. The participants were required to answer the same four questions whose difficulty was similar to those in the third part of IELTS test. The speech would be recorded and submitted. According to the rating criteria (The criteria for Part three in the IELTS speaking test was adopted), the raters would give the holistic score and analytic scores (include *Fluency and coherence, Lexical resource, Grammatical range and accuracy and Pronunciation*) as well.

In the current study, the MFRM adopted four-facets analysis, examinees' ability, rater's severity, rating criteria and rating scale, to answer the first two questions. The software for analysis was Minifac (Facets Student/Evaluation) Version No. 3.71.4.

The correlation analysis was done to answer the third research question. In this analysis, the correlation among all the four variables of spoken English (*fluency and coherence, vocabulary, grammar and Pronunciation*) and the weighting of analytic ratings to the holistic score were combined to proved evidence for the third question. The SPSS 23.0 was used to do this analysis.

4. Result and Discussion

4.1 MFRM Analysis Result



```
+------------------------------------------------+
|Measr|+examinee|-rater          |-criteria|Scale|
|-----+---------+----------------+---------+-----|
|   6 + *       +                +         + (9) |
|     | **      |                |         |     |
|     |         |                |         |     |
|     |         |                |         | --- |
|   5 +         +                +         +     |
|     |         |                |         |     |
|     | *       |                |         |     |
|     |         |                |         |     |
|   4 + *       +                +         +     |
|     |         |                |         |  7  |
|     | *       |                |         |     |
|     |         |                |         |     |
|   3 + *       +                +         +     |
|     | *       |                |         |     |
|     | **      |                |         | --- |
|     | *       |                |         |     |
|   2 +         +                +         +     |
|     | *       |                |         |     |
|     |         |                |         |     |
|     |         |                |         |  6  |
|   1 + **      +                +         +     |
|     | **      |                |         |     |
|     | *       |                | 3 4     |     |
|     | ***     |                |         | --- |
*   0 * **      * Rater1  Rater2 *         *     *
|     | *       |                | 1 2     |     |
|     | ***     |                |         |     |
|     |         |                | 5       |  5  |
|  -1 + *       +                +         +     |
|     |         |                |         |     |
|     |         |                |         | --- |
|     |         |                |         |     |
```

Figure 1. Wright Map from the Many-Facet Rating Scale Analysis

In the Wright map (Figure 1), the first column is the logit scale, based on which all the facets are positioned; the second column is the estimates of examinees' proficiency; the third column is the comparison of raters in terms of severity. Most severe rater takes an upper position, while the most lenient a lower position; the fourth column compares the difficulty of the five rating criteria; the last column is the nine-category rating scales. It can be easily seen that there is a wide range for the distribution of the examinees along the first logit scale, which means that the examinees can be divided into different levels quite well. The two raters are of the similar position, showing pretty good rating consistency.

Table 1. Examinee Measurement Report

Total score	Total count	Observed average	Fair(M) average	Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Examinee
64	10	6.40	6.40	2.22	.51	1.51	1.1	1.51	1.1	.49	.33	.33	01
58	10	5.80	5.81	.75	.46	1.11	.3	1.17	.4	.74	.30	.35	02
51	10	5.10	5.09	.55	.42	1.70	1.7	1.70	1.6	.39	-.01	.40	03
52	10	5.20	5.20	.38	.42	.61	-1.1	.6	-1.1	1.61	.67	.40	04
71	10	7.10	7.10	4.09	.52	1.78	1.5	1.78	1.5	.25	.20	.32	05
15	10	1.50	1.51	-9.04	.62	1.13	.5	1.14	.5	.70	-.17	.28	06
66	10	6.60	6.60	2.75	.51	.40	-1.5	.40	-1.6	1.60	.66	.33	07
77	10	7.70	7.70	5.69	.52	2.47	2.5	2.43	2.5	-.57	-.22	.32	08
78	10	7.80	7.80	5.96	.52	1.28	.7	1.25	.6	.71	.45	.32	09
73	10	7.30	7.29	4.62	.52	1.17	.5	1.18	.5	.83	.02	.32	10
54	10	5.40	5.41	-.03	.42	.75	-.5	.77	-.4	1.09	-.36	.39	11
59	10	5.90	5.91	.98	.48	.22	-2.2	.20	-2.3	1.70	.31	.34	12
51	10	5.10	5.09	-.55	.42	.87	-.2	.88	-.2	1.14	.19	.40	13
53	10	5.30	5.31	-.21	.42	1.01	.1	1.05	.2	1.13	.26	.39	14
56	10	5.60	5.61	.35	.44	.79	-.3	.79	-.3	1.21	.73	.37	15
65	10	6.50	6.50	2.49	.51	.97	.0	.96	.0	1.02	.47	.33	16
67	10	6.70	6.70	3.02	.52	.52	-1.1	.52	-1.1	1.48	.29	.32	17
55	10	5.50	5.51	.15	.43	1.07	.2	1.14	.4	.64	.73	.38	18
57	10	5.70	5.71	.54	.45	.23	-2.3	.21	-2.4	1.70	.69	.36	19
56	10	5.60	5.61	.35	.44	1.21	.5	1.26	.6	.64	.63	.37	20
77	10	7.70	7.70	5.69	.52	.49	-1.3	.48	-1.3	1.55	.37	.32	21
58	10	5.80	5.81	.75	.46	.50	-1.1	.52	-1.0	1.37	.67	.35	22
65	10	6.50	6.50	2.49	.51	.43	-1.4	.44	-1.4	1.56	.64	.33	23
59	10	5.90	5.91	.98	.48	.60	-.7	.65	-.6	1.30	.30	.34	24
69	10	6.90	6.90	3.55	.52	.21	-2.5	.20	-2.5	1.76	.43	.32	25
54	10	5.40	5.41	-.03	.42	1.57	1.3	1.59	1.3	.56	.45	.39	26
49	10	4.90	4.88	-.90	.42	.99	.0	.95	.0	1.2	.46	.40	27
62	10	6.20	6.20	1.71	.50	2.10	1.8	2.15	1.9	.28	-.08	.33	28
59.7	10	5.97	5.97	1.34	.48	.99	-.1	1	-.1		.34	Mean	
11.9	.0	1.19	1.19	2.83	.05	.56	1.3	.57	1.3		.30	SD(popul)	
12.2	.0	1.22	1.21	2.88	.05	.57	1.3	.58	1.3		.31	SD(sample)	

Model, Populn: RMSE .48 Adj (True) S.D. 2.79 Separation 5.81 Strata 8.08 Reliability .97

Model, Sample: RMSE .48 Adj (True) S.D. 2.84 Separation 5.92 Strata 8.22 Reliability .97

Model, Fixed (all same) chi-square: 758.8 d.f.: 27 significance (probability): .00

Model, Random (normal) chi-square: 26.2 d.f.: 26 significance (probability): .45

From Table 1, it can be seen that there was a wide spread of the examinees ability: from -9.04~5.96, more than 14 logits. The strata index for examinees' ability was 8.22, which meant that the examinees' ability could be divided into 8 classes, nearly corresponding to the 9 rating scales. And this also showed the good discriminancy of the 9 rating scales.

Table 2. Rater Measurement Report

Total	Total	Obsvd	Fair	Model		Inift		Outfit		Estim	Correlation		Exact	Agree	Rater
score	count	Average	Average	Meas ure	S.E.	MnSq	ZStd	MnSq	ZStd	Discrm	PtMea	PtExp	Obs%	exp%	
830	140	5.93	6.02	.09	.13	1.1	.8	1.11	.9	.92	.84	.85	35	42.9	1
841	140	6.01	6.09	-.09	.13	.87	-1.1	.88	-1	1.11	.86	.85	35	42.9	2
835.5	140	5.97	6.05	.00	.13	.98	-.2	1.0	.0		.85				Mean
5.5	.0	.04	.03	.09	.00	.11	1.0	.12	1.0		.01				S.D. populn
7.8	.0	.06	.05	.12	.00	.16	1.4	.16	1.4		.02				S.D. Sample

Model, Populn: RMSE .13 Adj (True) S.D. .00 Separation .00 Strata 3.33 Reliability(not inter-rater) .00

Model, Sample: RMSE .13 Adj (True) S.D. .00 Separation .00 Strata 3.33 Reliability(not inter-rater) .00

Model, Fixed (all same) Chi-square: 1.0 d.f.: 1 Significance (probability): .33

Inter-rater agreement opportunities:140 Exact agreements: 49 =35% Expected: 60.1 =42.9%

The focus of the current research was raters' rating reliability, which could be examined from two aspects: the severity and consistency. Table 2 showed that both of the two raters did exercise a similar level of severity which was supported by the separation index 0.33, that is, the two raters formed a single, homogeneous class. We should be cautious about the result, because there were only two raters. The intra-consistency of the individual rater could be supported by the Infit MnSq and Outfit MnSq. From Table 2, the Infit MnSq of the two raters were 1.10 and 0.87 separately, both within the range -2~2. So were the Outfit MnSq, 1.11 and 0.88. The above analysis suggested that the raters' rating were of high reliability.

Besides the aspect of raters, another focus of the research was the rating scale and the criteria. Bond & Fox (2007) argued that the indicators for assessing the effectiveness of rating scale categories are 1) the number of responses per category should be more than 10; 2) the responses frequency across categories be regular; 3) Model fit of rating scale should be <2.0; 4) The size of threshold increase is between 1.4 and 5.0; 5) There is a monotonic increase with category in terms of threshold.

Table 3. Category Measurement Report

Category	Score	Data count		Quality control			Rasch-andrich Thresholds		Expectation Measure at		Most Probable from	Rash thurstone threasholds
		Used	%	Avge Meas	Exp Meas	Outfit MnSq	Measure	S.D.	Category	-0.5		
1	0	5	2	-8.95	-9.16	1.2			(-10.07)		Low	Low
2	1	5	2	-9.13	-8.53	1.0	-8.99	.65	-7.03	-9.03	-8.99	-9.01
3												
4	2	21	8	-.46	-.54	1.1	-5.13	1.66	-3.15	-5.10	-5.13	-5.13
5	3	50	18	.03	.04	.9	-1.12	.25	-.70	-1.61	-1.12	-1.38
6	4	104	37	1.04	1.06	1.2	-.25	.17	1.20	.14	-.25	-.03
7	5	65	23	3.13	3.17	.9	2.50	.19	3.83	2.49	2.50	2.49
8	6	27	10	5.36	5.17	.8	5.16	.27	6.51	5.16	5.16	5.15
9	7	3	1	5.38	5.94	1.3	7.82	.61	8.95	7.98	7.82	7.88

Figure 4 presented all the five aspects mentioned above. The number of observations or responses in each rating scale is more than 10 except Categories 1, 2, 3 and 9; there is a peak in scale 6 with the next highest frequency observed in scales 7 and 5, which indicates that such unimodal distributions are generally unproblematic with regard to scale quality; the rating scale has an excellent model fit, that is, values of the Outfit MnSq were very close to the expected value of 1.0; From scales 4-8, there is a monotonic increase with category in terms of threshold, from -5.13~5.16. Because the scales 1~3 and 9 were less used, there were bigger values of thresholds, which meant lower scale quality.

From the above analysis, we can see that the nine rating scales are of high convergent and discriminant validity, that is, the 9 rating scales were useful and effective in evaluating the examinees' language ability.

Table 4. Criteria Measurement Report

Total Score	Total Count	Observed Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Z Std	Outfit MnSq	Z Std	Estim. Discrm	Correlation PtMea	Correlation PtExp	N Criteria
340	56	6.07	6.14	-.22	.20	.72	-1.5	.70	-1.7	1.33	.88	.85	1 HO
339	56	6.05	6.12	-.18	.20	.72	-1.5	.71	-1.6	1.28	.89	.85	2 FL
320	56	5.71	5.82	.56	.19	.86	-.7	.89	-.6	1.13	.85	.84	3 LX
320	56	5.71	5.82	.56	.19	1.19	1.0	1.28	1.5	.69	.80	.84	4 GR&AU
352	56	6.29	6.33	-.72	.21	1.44	2.0	1.42	1.9	.67	.80	.86	5 PN
334.2	56	5.97	6.05	.00	.20	.99	-.2	1.0	-.1		.84		Mean(count5)
12.5	.0	.22	.20	.49	.01	.28	1.5	.30	1.6		.04		S.D.(Populn)
13.9	.0	.25	.22	.55	.01	.32	1.6	.33	1.7		.05		S.D.(Sample)

Model, Populn: RMSE .20 Adj (True) S.D. .45 Separation 2.26 Strata 3.34 Reliability .84

Model, Sample: RMSE .20 Adj (True) S.D. .51 Separation 2.57 Strata 3.76 Reliability .87

Model, Fixed (all same) Chi-square: 30.2 d.f.: 4 Significance (probability): .00

Model, Random (normal) Chi-square: 3.5 d.f.: 3 Significance (probability): .31

For the estimation of criteria difficulty, it can be seen from Table 4, that *Pronunciation* is the least difficult, *Lexical resource, Grammatical range and accuracy* the most difficult, with *Holistic score* and *Fluency and coherence* lying in the middle. A careful examine on the Infit MnSq of the model showed that *Holistic score* and *Fluency and coherence* were slightly overfit because the value were -1.5 while Pronunciation was a little bit of misfit.

4.2 SPSS Analysis Result

Table 5. The Correlations of Rating Criteria

Descriptive Statistics

	Mean	Std. Deviation	N
HO	6.07	1.346	56
FL	6.05	1.394	56
LX	5.71	1.385	56
GR	5.71	1.371	56
PR	6.29	1.411	56

Correlations

		HO	FL	LX	GR	PR
HO	Pearson Correlation	1	.918**	.899**	.878**	.841**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	56	56	56	56	56
FL	Pearson Correlation	.918**	1	.856**	.836**	.760**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	56	56	56	56	56
LX	Pearson Correlation	.899**	.856**	1	.847**	.806**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	56	56	56	56	56
GR	Pearson Correlation	.878**	.836**	.847**	1	.786**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	56	56	56	56	56
PR	Pearson Correlation	.841**	.760**	.806**	.786**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	56	56	56	56	56

**. Correlation is significant at the 0.01 level (2-tailed).

As is shown in Table 5, the correlation analysis showed that the Holistic sore was highly correlated with *Fluency and coherency, Lexical resource, Grammatical range* and *accuracy*, and *Pronunciation* with correlation coefficient 0.918, 0.899, 0.878 and 0.841 respectively. The correlation coefficient between *Lexical resources* and *Fluency and coherence*, and *Grammatical range and accuracy* are 0.856 and 0.847. Even the lowest coefficient between *Fluency and coherence* and *Pronunciation*, 0.76 is more than 0.5. All the coefficient are more than 0.5, so there is a high convergent validity among all the five rating criteria. But on the other hand, it indicates low discriminant validity.

5. Conclusion

The construction validity of an analytic rating scale was investigated by the combination use of MFRM and SPSS correlation analysis. The MFRM provided a comprehensive analysis from the perspective of examinee, raters, rating criteria and rating scales, finding excellent consistency in rating severity and high rating reliability, and the validity of 9-scales to distinguish the examinees' speaking ability as well. The further look at the rating scale provided evidence for the convergent and discriminant validity of the rating scale. The correlation analysis told us that there was a good convergent validity but less discriminant validity because all of the five rating criteria were all highly correlated.

However, some problems did exist in the study: 1) the small sample in terms of both examinees and raters, may affect the generalizibility of the findings in this study; 2) the single speaking task was also a problem that made the research results less convincing. The future study could be improved by confirmatory factor analysis to support the validity of the rating scale and rating criteria.

References

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L., Lynch, B., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*, 238–57. http://dx.doi.org/10.1177/026553229501200206

Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language*

Learning, 34, 21–42. http://dx.doi.org/10.1111/j.1467-1770.1984.tb00350.x

Carroll, J. B. (1983). Psychometric theory and language testing. In Oller, J.W., Jr., editor, *Issues in language testing research,* Rowley, MA: Newbury House, 80–107.

Eckes, T. (2011). *Introduction to Many-Facet Rasch Measurement*. Frankfurt: Peter Lang.

Elder, C. (1993). How do subject specialists construe classroom language proficiency? *Language Testing, 10*, 235–54. http://dx.doi.org/10.1177/026553229301000303C

He, L. Zh., & Zhang, Jie. (2008). A Multifacet Rasch Analysis on the Validity of CET-SET. *Modern Foreign Languages*, *4*, 388-398.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*, 3–31. http://dx.doi.org/10.1191/0265532202lt218oa

Lee, Y.-W. (2005). *Dependability of scores for a new ESL speaking test: Evaluating prototype tasks.* (TOEFL Monograph Series, MS-28). Princeton, NJ: Educational Testing Service.

Lee. Y.-W., & Kantor, R. (2005). *Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes.* (TOEFL Monograph Series, MS-31). Princeton, NJ: Educational Testing Service.

Liu, J. D. (2005). The Multifacet Rasch analysis of Written Discourse Completion Test. *Modern Foreign Languages*，*2*, 157-169.

Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press.

Pollitt, A., & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing, 4*, 72–92. http://dx.doi.org/10.1177/026553228700400107

Sawaki Yasuyo. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language testing, 3*, 355-390. http://dx.doi.org/10.1177/0265532207077205

Wang, M.W., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research, 40*, 663-705. http://dx.doi.org/10.3102/00346543040005663

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-based Approach*. Basingstoke: Palgrave Macmillan.

Xi, X. (2003). *Investigating language performance on the graph description task in a semi-direct oral test.* Unpublished doctoral dissertation, University of California, Los Angeles.