International Journal of Educational Investigations 2016 (April), Vol.3, No.4: 1-12
Available online @ www.ijeionline.com ISSN: 2410-3446

Investigating Rater Effect on Iranian EFL Learners' Performance on Pragmatics Tests: An Application of Many-Facet Rasch Model

Reza Shahi[1]*

1. Dept. of English Language & Literature, Vali Asr University, Rafsanjan, Iran.

* Corresponding Author's Email: Reza.shahi411@gmail.com

Abstract **–** The current paper intends to investigate the impact of rater on test takers performance on pragmatic tests. In spite of the fact that different research studies have investigated the effect of construct irrelevant factors in other area of language teaching, few studies have investigated the impact of rater on test takers performance on pragmatic tests (e.g. Liu & Xie, 2014; Grabowski, 2008; Roever, 2008; Tajeddin & Alemi, 2014; Taguchi, 2011; Youn, 2007). Moreover, none of them has yet investigated the rater's effect on Iranian test takers performance on pragmatic tests. In this study, 55 senior students majoring in English literature filled out the Written Discourse Completion Tests (WDCT) along with the Oral Discourse Completion Test (ODCT). Two raters scored the WDCT and ODCT tests. Then, in order to examine the impact of the above mentioned factor, Many-facet Rasch measurement (MFRM) was implemented by using FACETS (Linacre, 2010). The results showed that there was a significant difference in rater severity in scoring pragmatic tests. Moreover, the bias analysis showed, raters showed significant bias across test takers and items. Although rater effect is inevitable, it can be minimized, if its nature is considered in advance.

Keywords**:** Pragmatic, Rater, Many facet ,WDCT ,and ODCT

1. INTRODUCTION

1.1. Pragmatics

Pragmatics can be defined as "the study of people's comprehension and production of linguistic action in context" (Kasper & Blum-Kulka, 1993, p .3). Pragmatics is always in the center of most models of the communicative competence that has been defined through the past century. Communicative competence is a term coined by Hymes (1972) to refer to "the aspect of our competence that enables us to convey and interpret messages and to negotiate meanings interpersonally within a specific context" (p.47). In a relative study, Canale and Swain (1980) differentiated between four components of communicative ability: grammatical, discourse, sociolinguistic and strategic competence. Based on Brown (1994), the first two components of Canale and Swain's communicative ability, reflected the use of the linguistics system, and the last two define the functional aspect of communication. Bachman (1990) coined the term of "language competence" which has two subcategories: organizational competence and pragmatic competence. The organizational competence deals with grammatical part, and pragmatic competence deals with sending and receiving massage by considering the sociocultural aspect of language.

1.2. Statement of the Problem

Investigating pragmatics is not a new trend in language studies but "only recently have issues related to pragmatic assessment attracted serious attention from second language acquisition (SLA) researchers" (Eslami, R, 2014, p.1). As Bachman (1990) noted, language testing and language teaching have reciprocal relationship. Therefore, language testing can serve as a source of information about the effectiveness of learning and teaching. Despite the increasing interest in pragmatic investigations, research on pragmatic testing still lag behind the assessment of other areas of language testing (Rover, 2006). Measuring learners' pragmatic knowledge is subjected to error like other fields of language testing. Some construct irrelevant factors, such as rater, may dramatically affect the results of pragmatic tests. So far, different studies have investigated the effect of different factors on pragmatic test scores in different contexts (e.g. Liu1 & Xie, 2014; Grabowski, 2008; and Tajeddin & Alemi, 2014). No study has yet investigated the impact of rater on Iranian test takers' performance.

2. LITERATURE REVIEW

In an attempt to develop testes to measure interlanguage pragmatics, Hudson, Detmer, and Brown (1995) developed a test battery that is widely used in assessing pragmatics. This test prototype consists of six types of testing methods: multiple-choice discourse completion test (DCT), open-ended DCT, oral DCT, role play, a self-assessment for the DCT, and a self-assessment for the role play. Following Hudson et al.'s (1995) projects, some researchers investigated the reliability and validity of the instruments that Hudson et al. (1995) developed in different target language teaching contexts (e.g. Brown, 2001; Hudson, 2001; Yamashita, 1996; Yoshitake, 1997). Besides, some researchers developed their own tests (e.g. Lui, 2006, Roever, 2001; Tada, 2005).

No matter how well pragmatic tests contextualized and stimulated real-world situations, they should still be scored by human raters, which have the negative effect by increasing costs and making them less practical (Tsutagawa, 2012). The different raters' effects can be summarized as (1) the severity effect, (2) the halo effect, (3) the central tendency effect, (4) inconsistency, and (5) the bias effect (Tajeddin & Alemi, 2014. p, 68). Research on rater effect in language testing in general has showed considerable degree of variability among raters. (Bachman, Lynch, Mason, 1995; Eckes, 2005; Kondo-Brown, 2002; Yang, 2010), however, few studies were conducted in pragmatic area to investigate the impact of rater.

Liu (2007), in a comparative study on native and nonnative English speakers' scoring in a WDCT, reported both four native and non-native raters that scored 38 students responses of a WDCT questioner were consistent in their overall ratings, but they were different in the severity.

Liu and Xie (2014) investigated the impact of rater on WDCT tests. They found significant differences between native and nonnative raters. Moreover, they found different rater bias across raters and rating trait and test takers. In another study, Youn (2007) examined rater bias in assessing Korean foreign language learners' performance. He found significant variation among raters. Also, he found significant bias across raters and other elements.

Taguchi (2011) studied differences between among NS raters who evaluated pragmatic performance of 48 Japanese subjects in two type of speech acts, found raters are different in their scoring. However, in a related study, Roever (2008) found no differences between raters' performance in his study.

3. METHODOLOGY

3.1. Purpose of the Study Research questions

Despite the increasing interest in pragmatic testing, few studies investigated the impact of rater on test takers performance. Therefore, this study intended to investigate the impact of rater on Iranian test takers' performance on DCTs.

This study addressed the following questions:

1. What are the main effects of test-taker ability, rater leniency, trait difficulty and item difficulty on test takers performance?
2. To what degree raters were different from each other in their level of leniency?
3. How reliably do the two raters reveal different degrees of leniency?
4. Do any of the raters show any particular bias pattern across different facts?
5. How appropriately are the five point rating scales functioning?

3.2. Participants

The study was conducted with 55 students studying at Vali-e-Asr University of Rafsanjan (VRU). The participants were composed of 40 female and 15 male English students. Their ages ranged between 17 to 24 years. They were selected conveniently from English translating and English literature students who had studied at least 2 semesters at VRU. Moreover, two raters were invited to score the test takers. The raters had at least six years language teaching experience.

3.3. Materials

Two types of measurement were used in this study: WDCT and ODCT. Each test consisted of 10 selective apology situations out of 12 situations which have been devised by Lui (2006). In each situation, test takers were asked to put themselves in each situation and write down what they say in such a situation.

> Example: You are a student. You forgot to do the assignment for your Human Resources course. When your teacher whom you have known for some years asks for your assignment, you apologize to your teacher.
>
> You_____ :
>
> (Liu, 2006: 197)

3.4. Procedure

First, the subjects were asked to do WDCT in 30 minutes. Then, two weeks later, test takers were invited separately to institute for ODCT administration. The DCT situations were read to the participants, and the participants were asked to respond orally to the situations. The performance of test takers was audio recorded. Finally, the transcriptions of the audio file were handed over to the raters. The raters used the rating scale that was developed by Hudson et al. (1995) to score test takers responses. They scored test takers performance based on their ability to use correct speech act (Speech act), appropriate wording and expressions (Expression), the amount of given information (Amount of info), and appropriateness of formality, directness, and politeness (Appropriateness).

4. RESULTS

4.1. Facets Preliminary Result

In order to report the effect of the mentioned factors, FACETS (Linacre, 2010) were used: A five-facet model was used to examine the effects of test-taker ability, rater leniency, test method difficulty, item difficulty and rating trait difficulty on students' performance on ODCT and WDCT.

One of the output files of facet is Facets Vertical Map or Rulers (Figure 1), which provides a graphical description of the variable (Linacere, 2010). The first column of the facet ruler is a measure which presents an equal interval representation of the test facet in logit. This provides a chance to interpret all of the facets in common. The second column represents the test takers' ability in logit. The test takers' measure in this study ranged from -.48 to 1.31 logit. Considering other statistics provided by the facet, a high reliability estimate of (.97) indicates facets ability to differentiate students with different levels of ability. A separation index of 5.34 is found for the test takers. A separation index is "a measure of the spread of the estimates relative to their precision" (Linacere, 2010. p, 197). And also significant fixed all same chi square (1551.1, p=.00) shows different levels of ability in test takers. Although all of the students are considered in the same level of ability, it suggests, there is a wide range of ability.

The second column shows raters' leniency in logit. As seen in the facet ruler, Rater 1 is more lenient than Rater 2, with logit measures of -.25for rater 1 and .25for rater 2. Considering other rater statistics, a high separation index of 19.32 with a high reliability index (1) and significant fixed all same chi square (374.1, p=.00) indicated Raters' different levels of leniency in this study.

The forth column shows test method difficulty in logit. WDCT with logit measures of -.09 and ODCT with logit measures of .09 are differentiated by facet. Although the logit spread of .18 is quite small, the relatively large separation index of 6.87 with a reliability index of .98 indicates that there are about five times method difficulty differences. And also significant fixed all same chi-square (48.2, p =.00) shows method differences in measuring students pragmatic.

The forth facet, items difficulty, can be seen in the fifth column. Item 2 with logit measures of .93 is the easiest item and item 10 with logit measures of .80 is the most difficult item. The High separation index of 14.64 and the high reliability index of 1 indicated the differences in items difficulty. Significant Chi square (1853.9, p=00), also proves the differences in items difficulty.

The last facet, rating trait, is presented in the sixth column. Appropriateness with logit measures .66 is the most difficult trait and speech with logit measures -.61 is the easiest trait. Significant chi squares (1195.5, p=00) and high separation and reliability which are respectively 12.36 and 99, indicated differences in difficulty level of rating traits in this study. And finally raw scores are presented in last coalmen along the same scale.

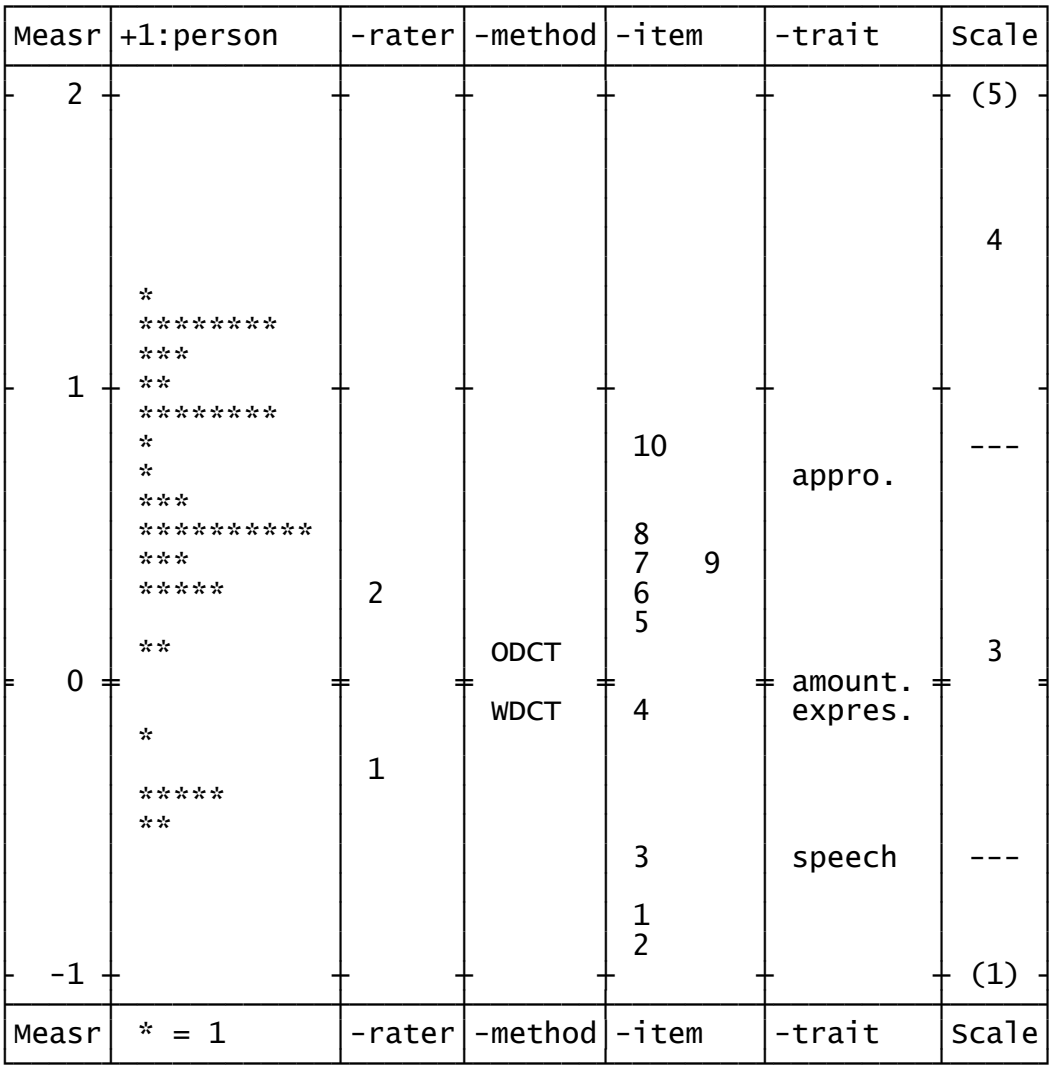

***Figure 1*: Facet vertical map**

4.2. Examining Central Tendency

Central tendency refers to the raters' tendency toward using middle scales which indicate raters' inability to differentiate students' ability through the test ((Myford & Wolfe, 2004). Table 1 presents the scale using of the raters in this study.

Table 1: ***Category Ctatistics***

DATA					QUALITY CONTROL				
Score	Category Total	Counts Used	%	Cum. %	Avge Meas	Exp. Meas	OUTFIT MnSq	Measure	S.E.
1	271	271	3%	3%	-.58	-.68	1.1		
2	1688	1688	19%	22%	-.16	-.16	1.1	-2.25	.06
3	2861	2861	33%	55%	.36	.38	.8	-.42	.03
4	2700	2700	31%	85%	.93	.94	1.0	.71	.03
5	1279	1279	15%	100%	1.51	1.48	1.0	1.96	.03

As frequency count (%) and cumulative frequency count (CUM %) show, categories in this study 1,2,3, 4 and 5 are used by the raters respectively 2% ,14%,34%,32 % and 18 % times of total category using. Category 3 is the category which is used more frequently and category 1 is used less frequently. Higher categories are used 50%of times and lower categories are used 16% of times.

4.3. Probability Curve

Figure 1 lists probability of occurrence for each category in this study. The Curve provides a graphical manifestation of the scales that were used. The horizontal axis is the test takers ability scale; the vertical axis gives the probability of rating by using of each scale. There is one curve for each category. Hill like curves are preferable. As figure shows, curve are somehow hill like, therefore, the category usage is reasonably well.

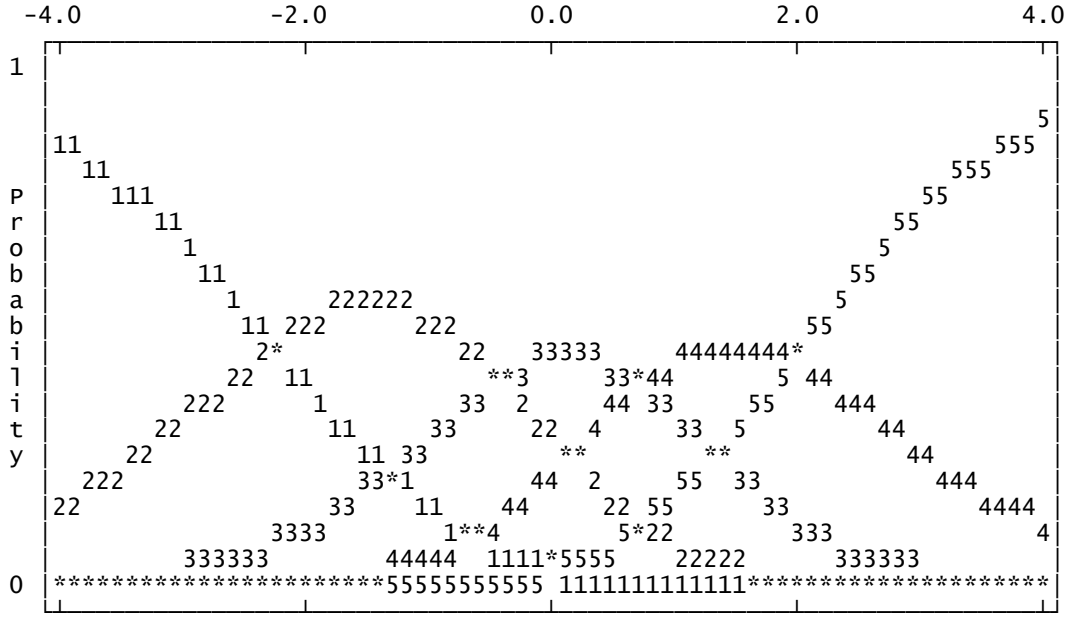

***Figure 2*: Probability curve**

4.4. Bias Analysis

The interaction of different facets especially raters and other facets are presented here. Raters may assign a score to a particular test taker that may be higher or lower than the expected model. Four bias analyses were conducted in this study.

4.4.1. Rater Bias across Test Method

The first bias analysis is related to the interaction of raters and test methods. Table 2 presents the bias statistics for the interaction of two facets. Considering control parameter places, all interactions bias measures are in expected range of ±1 logit and also in all interactions Z score are at the expected range of ±2. The consistency in bias pattern can be examined by misfit mean square. Considering the mean and standard deviation which are respectively 1 and 0.1, all mean squares are in expected zone (1. ±(0.1×2)=0.8~1.2) .

Table 2: ***Rater Bias across Test Method***

rater	method	Observed score	Expected score	average	Bias size	error	Z score	Infite
1	WDCT	7873	7863.88	.02	.01	.03	.24	1.1
1	ODCT	7588	7596.71	-.01	-0.1	.03	-.23	1
2	WDCT	6855	6846.07	-.01	-.01	.03	-.23	.9
2	ODCT	7588	7596.71	.00	.01	.03	.23	1

4.4.2. Rater Bias across Items

There are 20 total productions from 2 raters and 10 items within different methods .Four significant bias are found here. Rater one is harsher in scoring item one and two while, Rater two is more lenient in scoring the same items. Fit column also shows all items are in acceptable zone (1±(.2×2)= .6 -1.6). Here, only significant biases are presented.

Table 3: ***Rater Bias across Items***

rater	item	Observed score	Expected score	average	Bias size	error	Z score	Fit
1	1	1833	1785.69	.11	.19	.07	2.97	1.4
1	2	1850	1811.70	.09	.16	.07	2.45	.8
2	1	1602	1649.19	-.11	-.16	.06	-.2.77	1.1
2	2	1641	1679.17	-.09	-.13	.06	-2.27	.8

4.4.3. Rater Bias across Traits

Following table presents the interaction between rater and scoring trait. As can be seen in table, there are two significant bias out of 8 interaction which one of them is positive and the other one is negative bias. Both significant bias are related to appropriateness. A *z*-score below –2.0 shows that the rater consistently scores the trait more leniently compared to the way that particular rater scores other traits. On the other hand, a *z*-score greater than +2.0 suggests that the rater consistently scores the trait more harshly than others.

Table 3: ***Rater Bias across Traits***

rater	trait	Observed score	Expected score	average	Bias size	error	Z score	Fit
2	Appro.	3046	2987.18	.05	.08	0.4	2.17	1
1	speech	4334	4308.11	.02	.04	0.4	1	1.2
1	Amount.	3874	3846.54	.02	.04	0.4	1	1.2
1	Expres.	3946	3940.07	.01	.01	.04	.22	.8
2	Expres.	3556	3561.75	-.01	-.01	.04	-.21	.8
2	speech	3927	3952.61	-.02	-.03	.04	-.94	1
2	Amount.	3435	3462.32	-.02	-.04	.04	-.98	1
1	Appro.	3307	3365.87	-.05	-.08	.04	-2.12	1

Here also, the bias pattern is consistent. The standard deviation and means value of the fitness mean square were, 1.00 and .1 respectively. Here all of the values were in expected zone of .8 and 1.2.

4.4.4. Rater Bias across Test Takers by Considering Test Method

Table 4: ***Rater Bias across Test Takers***

rater	Test method	Test takers	Observed score	Expected score	Bias size	error	Z score	fit
2	WDCT	13	141	115.22	.92	.19	4.83	.9
2	ODCT	13	132	110.29	.78	.19	4.13	1.0
2	WDCT	8	156	138.69	.65	.20	3.25	1.5
2	WDCT	33	156	138.69	.65	.20	3.25	1.5
2	WDCT	39	156	139.21	.63	.20	3.16	1.5
2	WDCT	51	157	140.49	.63	.20	3.11	1.6

2	WDCT	21	156	139.72	.62	.20	3.07	1.5
2	ODCT	3	112	122.67	-.38	.19	-2.01	.5
2	ODCT	46	113	123.93	-.39	.19	-2.05	.5
2	WDCT	12	106	117.21	-.41	.19	-2.12	1.0
2	ODCT	28	112	124.19	-.44	.19	-2.29	.5
1	WDCT	21	139	153.12	-.53	.19	-2.78	.8
1	WDCT	39	138	152.64	-.54	.19	-2.87	.7
1	WDCT	8	136	152.17	-.60	.19	-3.16	.6
1	WDCT	33	136	152.17	-.60	.19	-3.16	.6
1	WDCT	51	136	153.83	-.66	.19	-3.51	.6
1	WDCT	13	106	129.28	-.84	.19	-4.33	1.0
1	ODCT	13	100	124.21	-.89	.20	-4.50	1.2

Although no rater bias over test method was reported in rater and method interactions, we can see here most of the rater bias across examines occurs in WDCT method and most of the biases is related to the rater 2 (11 out of 18). Negative Z value indicate more leniency and positive value shows more severity.

5. DISCUSSION

5.1. What are the main effects of test-taker ability, rater leniency, trait difficulty and item difficulty on test takers performance?

In terms of test-takers' ability, the facet analysis provided evidence that the test successfully separated the test takers into a wide range of ability levels. Thus, test takers were different from each other consistently and reliably. Considering the test type, significant chi square and high reliability index, showed test methods are different from each other consistently. In addition, facet analysis showed the rating treats were different from each other consistently. The finding of this part of the study is in line with Youn (2007) who found the significant difference between these elements. Lui (2014) also found the elements within in different facts in WDCT test were significantly different from each other and have significant effect on test takers performance. In addition, this is in line with the finding of Gorbowki (2008) that found different variables such as test takers ability, item difficulty have significant effect on test takers performance.

5.2. To what extent raters were different from each other in their level of leniency?

Data analysis indicated that the raters differed significantly in their level of severity. Rater 1 was more lenient than Rater 2, with logit measures of -.25for rater 1 and .25for rater 2.

The rater finding showed some differences between raters in most of the previous studies (e.g. Brown &Ahn, 2010; Grobowski, 2008; Lui, 2014; and Youn, 2007). However, Rover (2006) didn't find any differences between raters performance in DCT test. Tajeddin and Alemi (2014), suggested rater training in order to minimize the rater effect.

5.3. How reliably do the two raters reveal different degrees of leniency?

Following FACETS result which provides a reliability coefficient for each facet to show how reliably the facet distinguishes the elements within each facet. In rater's analysis, the reliability indicates how reliably the raters were separated into different degrees of severity (Youn, 2007). The reliability coefficient was 1.00. Therefore the analysis quite reliably reveals the different degrees of leniency among raters.

5.4. Do any of the raters showed any particular bias pattern across different facts?

The bias analysis provides *z*-scores, which represent the degree of differences between the expected scores and the observed scores. Also, the *z*-scores, either above +2.00 or below -2.00, indicate significance bias. Considering bias analysis of raters across test method, the bias between the two raters and the two test method was non-significant. However, raters across items bias analysis showed four significant bias that rater one was harsher in his scoring. Moreover, there wasn′t any significant bias pattern in raters and traits interaction.

Although any rater bias over test method and rating trait wasn't reported in rater and method interactions, some significant rater bias across test takers in different method were reported. Most of the rater bias across examines occurs in WDCT method and most of the biases was related to the rater 2 (11 out of 18). Negative Z value indicate more leniency and positive value shows more severity. However, there was no predictable pattern in their bias pattern. Totally speaking, rater 2 was harsher in scoring test takers.

5.5. How appropriately are the five-point rating scales functioning?

The analysis of scale functionality of the scales indicated that the rating scales seemed to be functioning properly. Considering the percentage of the scale usage, although the middle scale had the highest percentage, the total distribution of scale usage was acceptable. The finding of this part is in line with other studies (e.g. Lui, 2014; Brown & Ahn, 2010; Gorbowski, 2008). They found the five point scales functions well in pragmatic testing

6. CONCLUSION

This study aimed to detect the effect of rater on students' performance. Two test format, Written Discourse Completion Tests (WDCT) and Oral Discourse Completion Test (ODCT), were used in this study. The WDCT and ODCT was scored by two raters. MFACT was used to investigate the differences of the elements within each facet which was included: test takers, rater, and items in each separated test. Facet analysis showed elements within different

variables were different from each other significantly. Moreover, variables which were same in two test types showed some differences. Based on the fact ruler and student measures table, students had significant difference in their performance on the tests. It was indicator of different level of ability which showed tests successfully differentiated students with different ability and high reliability and separation index showed they are different from each other consistently.

Item characteristic table and its ruler map showed items had different level of difficulty in both tests. Raters in WDCT and ODCT showed significant differences in their leniency as significant chi square showed in rater measurement table. Bias analysis revealed raters did not showed and bias across test method and rating traits. However, they showed significant bias across test takers and items. The impact of rater as a source of variance is not evitable, but it can be minimized by rater training.

REFERENCES

Bachman, L.F., Lynch, B.K., Mason, M., 1995. Investigating variability in tasks and rater judgments in a performance test of foreign language speaking.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Brown, A., 1995. The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing* 12, 1–15.

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

Hudson, T. (2001). Indicators for pragmatic instruction. In K. R. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 283-300). Cambridge: Cambridge University.

Hudson, T., Brown, J. D., & Detmer, E. (1995). *Developing prototypic measures of cross-cultural pragmatics* (Vol. 7). Natl Foreign Lg Resource Ctr.

Hinkel, E. (1997). Appropriateness of Advice: DCT and Multiple Choice Data1. *Applied linguistics*, *18*(1), 1-26.

Kondo-Brown, K., (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. Language Testing 19, 3–31.

Liu, J. (2004). *Measuring interlanguage pragmatic knowledge of Chinese EFL learners* (Doctoral dissertation, City University of Hong Kong).

Liu, J., & Xie, L. (2014). Examining Rater Effects in a WDCT Pragmatics Test. *Iranian Journal of Language Testing*, *4*(1).

Liu, J. (2007). Comparing native and nonnative speakers' scoring in an interlanguage pragmatics test. *Modern Foreign Languages*, *30*(4), 395-404.

Grabowski, K. C. (2008). Investigating the construct validity of a performance test designed to measure grammatical and pragmatic knowledge. *SPAAN FELLOW, 1001*, 131.

Rintell, E.M., Mitchell, C.J., 1989. Studying requests and apologies: an inquiry into method. In: Blum-Kulka, S., House, J., Kasper, G. (Eds.), Cross-cultural Pragmatics: Requests and Apologies. Ablex, Norwood, NJ, pp. 248–272.

Rose, K. R. (1994). On the Validity of Discourse Completion Tests in Non-Western Contexts1. *Applied Linguistics*, *15*(1), 1-14.

Roever, C. (2001). *A web-based test of interlanguage pragmalinguistic knowledge: Speech acts, routines, and implicatures*. Unpublished doctoral dissertation.

Tada, M. (2005). *Assessment of ESL pragmatic production and perception using video prompts*. Unpublished doctoral dissertation, Temple University, Japan.

Taguchi, N. (2011). Teaching pragmatics: Trends and issues. *Annual Review of Applied Linguistics, 31*, 289-310.

Tsutagawa, F. (2012). Future directions in pragmatics assessment. *Working Papers in TESOL & Applied Linguistics*, *12*(2), 43-45.

Tajeddin, Z., & Alemi, M. (2014). Pragmatic rater training: Does it affect non-native L2 teachers' rating accuracy and bias. *Iranian Journal of Language Testing*, *4*(1), 66-83.

Sasaki, M. (1998). Investigating EFL students' production of speech acts: A comparison of production questionnaires and role plays. *Journal of pragmatics*, *30*(4), 457-484.

Varghese, M., & Billmyer, K. (1996). Investigating the Structure of Discourse Completion Tests. *Working Papers in Educational Linguistics*, *12*(1), 39-58.

Yamashita, S.O., 1996. Comparing six cross-cultural pragmatics measures. Unpublished doctoral dissertation, Temple University, Philadelphia, PA

Yamashita, S. O. (1997). Self-Assessment and Role Play Methods of Measuring Cross-Cultural Pragmatics. *Pragmatics and Language Learning*, *8*, 129-162.

Yang, Rui. (2010). A Many-facet Rasch Analysis of Rater Effects on an Oral English Proficiency Test. (PhD), Purdue University, West Lafayette, Indiana, USA.