Tabaran Institute of Higher Education
Iranian Journal of Language Testing
Vol. 4, No. 1, March 2014

ISSN 2251-7324

Examining Rater Effects in a WDCT Pragmatics Test

Jianda Liu[1], Lijun Xie[2]

Abstract:

Written Discourse Completion Task (WDCT) has been used in pragmatics tests to measure EFL learners' interlanguage pragmatic knowledge. In a WDCT, the students give their responses to situations designed to elicit certain pragmatic functions, so human raters are required to rate the students' performance. When decisions are made based upon such ratings, it is essential that the assigned ratings are accurate and fair. As a result, efforts should be taken to minimize the impact of rater inaccuracy or bias on ratings. This paper reports a study of rater effects in a WDCT pragmatics test. Based on the Myford& Wolfe (2003; 2004) model and corresponding retrospective interviews, four types of rater effects were investigated and discussed quantitatively and qualitatively: leniency/severity, central tendency, halo effect, and differential leniency/severity. Results revealed significant differences in terms of rating severity, with a general tendency towards severity. Though the raters could effectively and consistently employ the rating scales in their ratings, some of them showed certain degrees of halo effect. Most raters were also found to exhibit certain bias across both traits and test takers. Possible reasons behind the rater effects were analyzed. Finally suggestions were raised for rating training.

Keywords:*WDCT, rater effects, Pragmatics test, rater training*

1. Introduction

Pragmatic competence can be broadly defined as the ability to use language appropriately in a social context (Taguchi, 2009). It can also be understood as the knowledge of the linguistic resources available in a given language for realizing particular illocutions, knowledge of the sequential aspects of speech acts, and finally, knowledge of the appropriate contextual use of the particular language's linguistic resources (Barron, 2003). As a domain within L2 studies, pragmatics is usually referred to as interlanguage pragmatics (ILP). Interlanguage pragmatic competence can thus be defined as the nonnative speaker's knowledge of a pragmatic system and knowledge of its appropriate use (Kasper, 1998). As part of the communicative language ability defined by Bachman (1990), interlanguage pragmatic competence has attracted more and more attention in language testing. Such an interest yields various measures to test the ESL learners' interlanguage pragmatic knowledge. To date, at least six measures have been

[1] *Guangdong University of Foreign Studies, China, Email: jackliu@gdufs.edu.cn*
[2] *No. 2, Middle School of Xiamen, Fujian, China, Email: 120867338@qq.com*

developed (Brown, 2001; Hudson, Detmer, & Brown, 1995), among which the Written Discourse Completion Task (WDCT) is often used by researchers for data collection and testing purposes. In the WDCT test, test takers are required to provide a response that they think appropriate in a given context. Their pragmatic ability is estimated by means of assessing their responses by human raters according to certain rubrics.

When decisions are made based upon such ratings, it is essential that the assigned ratings be accurate and fair. However, unavoidably human raters may introduce errors into the final scores for different reasons, such as unfamiliarity with or inadequate training towards the rating scale, fatigue or lapses in attention, deficiencies in some areas of content knowledge, or personal beliefs that conflict with the values espoused by the scoring rubric (Wolfe & Chiu, 1997). As a result, test authorities try their best to employ different means to improve the rating reliability, such as rater selection, training, and various monitoring procedures. These measures help to improve the rating reliability, but idiosyncrasies still exist in the behaviors of raters though great efforts are taken to minimize inaccuracy and bias in ratings (Bonk & Ockey, 2003; Elder et al., 2005; Lumley & McNamara, 1995; Wolfe, 2004).

Pragmatics testing is at its initial development stage. Currently the fact is that there are more questions about assessing pragmatics than there are answers (Cohen, 2008). WDCT is widely used in the field of pragmatics and many related studies have been reported, mainly because of its simplicity of use and high degree of control over variables (Brown, 2001; Golato, 2003). However, few research has been done to investigate its use in pragmatics assessment. One of the concerns on its use in pragmatics testing is its reliability which relies largely on the performance of the raters. Systematic patterns in idiosyncratic behaviors of the raters are normally termed as rater effect or bias (Wolfe, 2004). Rater effect thus becomes one of the major concerns in using WDCT in pragmatics tests. This study attempted to explore whether, how and why some common patterns of rater effects might exist in a WDCT pragmatics test.

2. Literature review

Methods for measuring ILP knowledge were mostly derived from the ILP data collection measures. WDCTs are written questionnaires including a number of brief situational descriptions. Respondents are asked to provide a response that they think is appropriate in the given context. Many studies on WDCT as a data elicitation tool have been reported (e.g., Billmyer & Varghese, 2000; Rose, 1994), while few research on WDCT as a testing method has been conducted. So far, research on WDCT as a testing instrument has centered on the validation of the test method itself. Hudson et al. (1995) found that, in WDCT, NSs and NNSs basically used similar strategies although their responses varied according to different speech acts and situations. Yamashita (1996), Roever (2005; 2006), and Liu (2006) all revealed that WDCT was basically reliable and valid. Studies by Yoshitake-Strain (1997) and Enochs and Yoshitake-Strain (1999), however, showed that WDCT was not highly reliable or valid in assessing pragmatic competence when administered to Japanese university EFL students.

Different results of the existing studies indicate the necessity of more such studies. One of the difficulties in applying WDCT in pragmatics tests is that raters are required to score the responses given by the test takers. In a WDCT pragmatics test, the use of ratings assumes that the raters are reasonably objective and accurate, so raters are required to objectively reach a conclusion about the test takers' performance. However, it is not an easy task, for, in reality, in addition to the differences in using the rating rubrics, raters' memories are quite fallible, and raters subscribe to their own sets of likes, dislikes, and expectations about people, which may or may not be valid (Kumar, 2005). Rater variation and bias may affect the reliability of the ratings. Rater bias or effect refers to the systematic deviations between the “true” rating a test taker deserves and the actual rating assigned (Myford & Wolfe, 2003; Scullen, Mount, & Goff, 2000). The deviations can manifest themselves in various forms, such as the degree to which raters comply with the scoring rubrics, the degree to which their ratings are consistent across examinees, scoring criteria, and performance tasks, etc. (Eckes, 2008; Lumley & Brown, 2005).

Research on rater effect in language performance assessments has provided ample evidence for a considerable degree of variability among raters. Raters were found to show variations in severity (Bachman et al., 1995; Kondo-Brown, 2002; Eckes, 2005; Yang, 2010). Rating variation was also detected between raters and domains (Gyagenda& Engelhard, 1998), and in raters’ views on the importance of the various criteria (Eckes, 2008). Some researchers also tried to explore the possible sources of the score variance. Hsieh (2011) found that raters’ experience with accented speech, perceptions of accent was an important rating criterion, and approaches to rating (i.e. analytical or global) had important bearings on raters’ judgments. Wiseman (2012) examined the decision-making behaviors of raters when scoring essays written by second language learners. Results suggested that rater background might have contributed to rater expectations that might explain individual differences in the application of the performance criteria of the rubrics when rating essays.

Compared to the studies on rater effect in language performance assessment, few has been done to investigate the rater effect in interlanguage pragmatics assessment. Liu (2007) reported a comparative study on native and nonnative English speakers’ scoring in a WDCT interlanguage pragmatics test which contains 12 request situations. Eight raters were invited to rate the responses from 38 participants according to the analytic rating rubrics developed by Hudson et al. (1995), with four rating dimensions: speech act, amount of information, expression, and appropriateness. Results demonstrated that both native and nonnative English speakers, though fairly consistent in their overall ratings, differed strongly in the severity. Nonnative English speakers (NNSs) were found to be more lenient than the native English speakers (NSs). Step disordering was found among both the NSs and the NNSs in the dimension of speech act, but in the dimensions of amount of information and appropriateness such disordering was detected only in NNSs. Walters (2007) applied conversation analysis (CA) to detect rater variations in a test of ESL oral pragmatic competence. Two CA-trained raters (one NS and one NNS) rated the responses based on a four-point holistic rating scale. After all responses had been rated, the raters held a series of dialogues regarding differentially rated performances in order to resolve differences in scoring between the raters. The results showed that different scoring decisions were made by the two raters due to different interpretations of the examinees' performance. The NS rater sometimes relied on his knowledge of normative patterns, while the NNS rater was sometimes influenced by the examinee's fluency and clear pronunciation. Taguchi (2011) explored variability among NS raters who evaluated pragmatic performance of learners of English as a foreign language. Four English NSs of mixed cultural backgrounds assessed the appropriateness of two types of speech acts (requests and opinions) produced by 48 Japanese EFL students. Norms and the reasoning behind the raters’ assessment practice were investigated through individual introspective verbal interviews. Results revealed divergent focus of the four raters when evaluating appropriateness of the speech acts. Some raters were more focused on linguistics forms, while others based their scoring decision on non-linguistics aspects such as the use of positive/negative politeness strategies and semantic moves as well as the content of speech. Some raters even incorporated additional, unique features that they felt were salient into the evaluation criteria. Even when focused on the same dimension, the raters differed in their degree of acceptance. Some raters also based some of their assessment decisions on their own personal experiences. Youn (2007) investigated whether various factors, including test types, speech acts, groups of candidate, and test items, affected raters' assessment of the pragmatic competence of KFL learners in terms of request and apology in the format of WDCT, Language Lab, and Role-play. Results indicated that all three raters showed different degrees of severity in their ratings, depending on the test type and speech act. Additionally, each rater displayed unique bias patterns within the interactions.

3. Methodology

3.1. Research questions

Previous research concerning rater variation in pragmatics tests has identified several ways that raters may introduce errors into examinee scores. However, very few studies have examined the actual rating process, especially the raters' mental process. In fact, many questions remain unanswered. The questions raised by Taguchi (2011) deserve immediate attention:

How do raters interpret and internalize descriptions of rating rubrics? Do they bring their own criteria in determining appropriateness of pragmatic behaviors? Do they prioritize one dimension of pragmatic appropriateness over others, and is there variation in their orientation?(Taguchi, 2011: 455)

Based on a model proposed by Myford and Wolfe (2003, 2004), this study quantitatively analyzed the possible rater effects in a WDCT pragmatics test, and explored the possible reasons behind the rater effects though introspective interviews. Specifically, this study addressed the following questions:

a) Do raters differ in the levels of severity in their WDCT ratings?
b) Do raters effectively and consistently employ the rating scales in their WDCT ratings?
c) Do raters efficiently differentiate between traits, that is, do raters show any evidence of halo effect?
d) Do raters exhibit bias in their WDCT ratings?

3.2. Test materials

The WDCT test paper used in this study was adopted from Liu (2006) which contains 12 apology situations in which test takers are required to write down what they think would be an appropriate response for each situation. For 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)

3.3. Examinees and raters

The WDCT test was administered to 38 (15 males and 23 females) Chinese EFL university students aged from 19 to 21. All of them were students majoring in English from tertiary universities in China. They had studied English for about 10 years and their English proficiency could be rated basically as the upper-intermediate level （FCE according to the

University of Cambridge ESOL Examinations）. The raters were 6 university EFL teachers. Four of them were Chinese EFL teachers with a relatively high English proficiency, while the other two raters were native speakers of English teaching English in China. Although NS raters and NNS raters were found to show differences in rating WDCT tests (Liu, 2007; Youn, 2007), they were treated as the same rating team in this study, considering the fact that such differences existed among both the NS raters and the NNS raters. To further explore the inner thoughts of the raters while rating, retrospective interviews were conducted with the 6 raters. All the interviews were recorded and coded.

3.4. Procedure

The test was conducted in a classroom. The writer first explained to the students what the test was intended for and how the students were supposed to answer the items. To ensure the authenticity of the data collected, the writer embedded this test as part of the tasks of a teaching unit in a major credit course named Communicative English for EFL learners. Though time limit was not set, the test lasted for about 35 minutes.

Then, 6 teachers were invited to rate the responses given by the students. The rating rubrics were based on the rating scale developed by Hudson et al. (1995) which required raters to assess the students' responses on a 5-point Likert scale in terms of their ability to use the correct speech act (Speech act) and appropriate expressions and wording (Expression), the amount of information given (Amount of info), and levels of formality, directness, and politeness (Appropriateness). To avoid any effect on ratings due to poor handwriting, the students' responses were entered into computer without any changes. The typewritten scripts were ordered alphabetically according to the test takers' surnames and then presented to the raters. The raters were given clear directions as to how the test papers should be rated and had preliminary training on the rating. First, a presentation was given to the raters about the nature of the five common rater errors (Saal, Downey, & Lahey, 1980) with an aim to sensitize the raters to the type of errors they might commit. Second, a training manual based on the one developed by Hudson et al. (1995) was given to the raters. In the training manual, the speech act of apologies and rating criteria were explained in detail. Then, the raters were required to rate three samples, after which a discussion was held. The ratings of the raters for each item were compared and discussed. When a discrepancy occurred, opinions were exchanged and a general consensus was reached. The rating was done in the raters' free time and lasted for about a month.

4. Results

4.1. Rater reports

The results reported in this paper were computed through the software FACETS (Linacre, 2012). Table 1 summarized how the raters used the scale categories across all trait scales. The significant chi-square (χ^2=1883.2, df=5, p<.01) showed that the raters did not exercise the same level of severity in their ratings. The rater separation ratio (19.59) indicated that the differences between rater severities were almost 20 times greater than the error with which these severities were measured. The rater separation index (26.45, using the formula (4G + 1) / 3, where G is the rater separation ratio) (Myford & Wolfe, 2004) suggested that there were about 27 statistically distinct strata of rater severity in this sample of raters.

Results revealed various severities of the 6 raters, with a difference of 1.85 logits between the most severe rater (Rater A, 0.82 logits) and the most lenient rater (Rater E, -1.03

logits). The mean severity of the 6 raters is 0.00 logits with a standard deviation equaling to 0.61. In this study, 4 raters' (A, B, C, and D) fair averages were below the mean (2.89).

Table 1 Rater measurement report

Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Raters
2.4	2.41	.82	.03	1.00	.0	1.00	.0	A
2.7	2.72	.30	.03	.66	-9.0	.66	-9.0	C
2.8	2.86	.06	.03	.67	-9.0	.67	-9.0	D
2.9	2.88	.03	.03	1.50	9.0	1.51	9.0	B
3.0	3.00	-.18	.03	.94	-1.8	.94	-1.9	F
3.4	3.46	-1.03	.03	1.28	7.8	1.27	7.6	E
2.9	2.89	.00	.03	1.01	-.5	1.01	-.6	Mean
.3	.35	.61	.00	.33	7.8	.34	7.8	S.D
Separation 19.59; Reliability 1.00; Chi-square 1883.2; Significance: .00								

4.2. Trait reports

Results of the trait analysis were reported in Table 2. Different difficulties (0.53 logits difference) were found among the four traits. Appropriateness and Expressions were the most difficult traits (0.15 logits) and Speech Act was the easiest one (-0.38 logits). The significant chi-square value (χ^2=297.1, *df*=3, *p*<.01) rejected the null hypothesis that all traits were of the same calibrated level of difficulty, indicating that at least two traits were significantly different in terms of their difficulty.

The trait separation ratio of 9.97 indicated that the spread of the trait difficulty measures was about 10 times larger than the precision of those measures. The trait separation index of 13.6 signaled that there were nearly 14 statistically distinct strata of trait difficulty. And the high degree of trait separation reliability of 0.99 suggested that the raters could reliably distinguish among the traits. For all the traits, their infit mean squares were within the acceptable range.

Table 2 Trait measurement report

Obsvd Average	Fair-M Avrage	Measure	Infit MnSq	ZStd	Outfit MnSq	ZStd	Traits
2.8	2.81	.15	1.13	4.8	1.14	4.9	Appro.
2.8	2.81	.15	.88	-4.5	.89	-4.4	Expres.
2.8	2.85	.08	1.06	2.1	1.06	2.2	Amount
3.1	3.11	-.38	.94	-2.1	.94	-2.2	Speech
2.9	2.89	.00	1.00	.1	1.01	.1	Mean
.1	.14	.26	.11	4.2	.11	4.3	S.D
Separation 9.97; Reliability .99; Chi-square 297.1; df 3; Significance .00							

4.3. Use of the categories of the rating scale

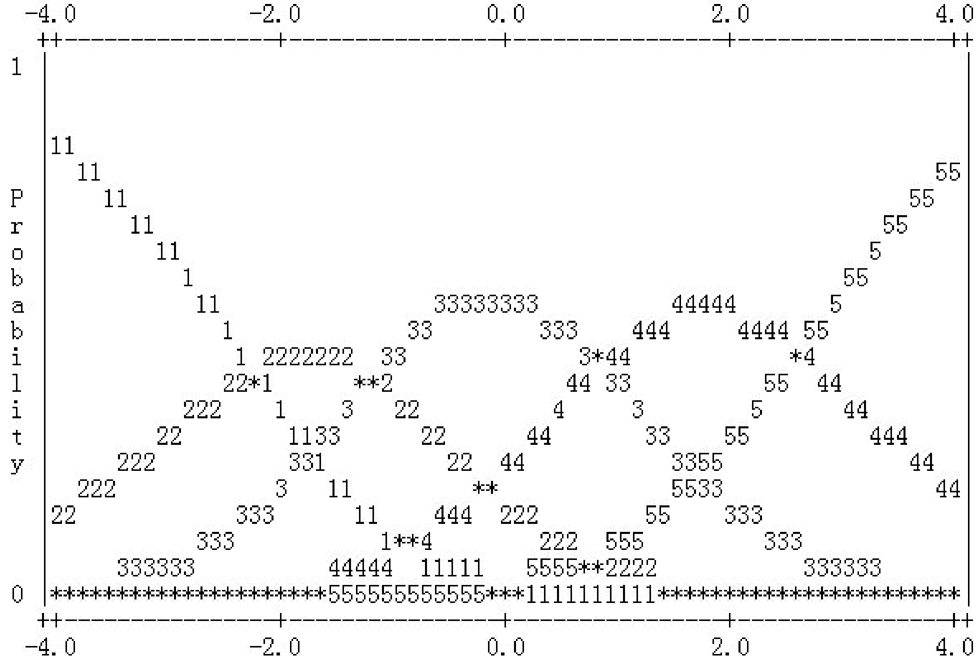
When a rater overuses the middle categories of a rating scale, this rater may exercise the "central tendency effect", indicating that the rater is most probably unable to differentiate among examinee performance levels along the entire performance continuum (Myford &Wolfe, 2004). Table 3 showed the scale category statistics. From the frequency count (column 2) and percentage of ratings (column 3) that the raters assigned in each rating scale category, we can see Category 3 was used the most (45%), followed by Category 2 (23%) and Category 4 (21%). Category 5 was the least used (3%). The raters as a group used the lower rating scale categories (1 and 2) 31% of the times and the higher rating scale categories (4 and 5) 24% of the times.

Both the step calibration and the probability curves (Figure 1) signified that the rating scale adopted in this study functioned reasonably well.

Table 3 Category statistics

DATA				QUALITY CONTROL			STEP CALIBRATIONS	
Category Counts				AvgeMeas	Exp.	OUTFIT		
Score	Used	%	Cum. %		Meas	MnSq	Measure	S.E.
1	855	8%	8%	-1.49	-1.46	1.0		
2	2555	23%	31%	-.72	-.83	1.1	-2.24	.04
3	4961	45%	76%	-.32	-.25	1.0	-1.20	.02
4	2245	21%	97%	.37	.35	1.0	.84	.03
5	328	3&	100%	1.27	1.01	.8	2.60	.06

Figure 1 Probability curves



```
   -4.0             -2.0               0.0               2.0              4.0
    ++----------------+-----------------+-----------------+----------------++
  1 |                                                                       |
    |                                                                       |
    |                                                                       |
    |11                                                                     |
    |  11                                                                 55|
  P |    11                                                             55  |
  r |      11                                                         55    |
  o |        11                                                     5       |
  b |          1                                                  55        |
  a |           11                 33333333          44444       5          |
  b |             1              33        333    444     4444 55           |
  i |              1 2222222   33             3*44            *4            |
  l |             22*1       **2             44 33          55  44          |
  i |          222    1    3    22          4     3        5      44        |
  t |        22        1133       22      44       33    55         444     |
  y |     222          331          22  44           3355              44   |
    |  222            3   11          **             5533                44 |
    |22            333      11     444  222        55    333                |
    |           333            1**4        222  555         333             |
    |     333333           44444  11111   5555**2222           333333       |
  0 |**********************555555555555***1111111111************************|
    ++----------------+-----------------+-----------------+----------------++
   -4.0             -2.0               0.0               2.0              4.0
```

4.4. Bias analysis

Differential severity/leniency occurs when a rater tends to assign ratings to a particular group of examinees that are, on average, lower/higher than the measurement model would expect for that group, given other raters' ratings of the group (Myford & Wolfe, 2004). To investigate whether raters maintained a uniformed level of severity in their ratings, two bias analyses were conducted.

4.4.1 Rater bias across examinees

The first bias analysis was meant for interaction between the raters and the examinees. There were 228 total productions from 6 raters and 38 candidates. The analysis revealed 87 (38%) significant bias interactions between raters and examinees, among which 46 (20%) interactions tended towards unexpected severity and 41 (18%) tended towards unexpected leniency. All 6 raters had significant bias interactions for the examinees, with a mean of 14.5. Rater D had the minimum significant interactions (7), while Rater B had the maximum (23). Among the 38 examinees, 33 examinees had significant bias interactions with the raters, ranging from 1 interaction (Examinee 12 and Examinee 15) to 5 interactions (Examinee 9). Four raters (A, B, D, E) produced altogether 11 misfitting interactions.

To investigate whether there were systematic patterns in rater-examinee interactions, a closer examination of the bias interactions was performed. Based on the method used by Schaefer (2008), Table 4 showed the bias patterns for raters across examinees' ability levels. The examinees were divided into five groups according to their ability logits, with 0.4 logits (mean +1 SD) as a dividing line for groups, from the highest ability level, 1.6 logits, to the lowest, -2.16 logits. The second row of Table 4 was the number of examinees in each ability group. The bias interactions for all 6 raters were divided into severe and lenient ratings. Among the 87 significant bias interactions, there were slightly more severe ratings (46) than lenient ratings (41). Almost half of the bias interactions (51%, 44 out of 87) occurred around one standard deviation from the mean, between -1.04 and 0.4 logits. Fifty-three (61%) bias interactions happened on the lower ability examinees (ability measure below 0.00 logits), indicating that raters were more likely to show bias towards lower ability examinees than higher ability examinees. Bias interactions for lower ability examinees were slightly more likely to be lenient than severe: 24 severe to 29 lenient. And the number of bias interactions for higher ability examinees was equal: 17 severe to 17 lenient. Schaefer (2008) found that examinees at extreme ends of the scale tended to attract more bias interactions, which was echoed in this study in that Examinee 1, located at the extremely high place, received 3 biases and Examinee 9, located at the extremely low place, got 4 biases.

Table 4 Frequency of significant Rater-Examinee bias

Testee	0.88-	0.12-	-0.64-	-0.65- -	-1.41- -	
logits	1.60	0.87	0.11	1.40	2.16	Total
N. of testees	1	10	16	7	4	38
Severe/lenient	S/L	S/L	S/L	S/L	S/L	S/L
Rater A	0/1	1/5	5/3	4/0	0/1	10/10
Rater B		2/1	3/7	3/4	3/0	11/12
Rater C		2/2	1/4	2/2	0/2	5/10
Rater D	1/0	2/0	2/0		0/2	5/2
Rater E	1/0	3/1	3/3	0/1	0/2	7/7
Rater F		0/2	2/2	0/1	0/1	2/6
Total	2/1	10/11	16/19	9/8	3/8	40/47

4.4.2. Rater bias across traits

All six raters had significant bias interactions with traits. There were altogether 15 (62.5%) significant bias interactions out of the total 24 interactions, of which 9 had negative bias sizes (showing severity) and 6 had positive ones (showing leniency). Among the traits, there were

5 significant bias interactions for Speech Act (3 negative and 2 positive), 3 for Expressions (2 negative and 1 positive), 3 for Amount of info (2 negative and 1 positive), and 4 for Appropriateness (2 negative and 2 positive). The number of significant biases for traits shown by individual raters ranged from 1 to 3. All raters who showed significant bias interactions had bias for the trait Speech Act except Rater F. Raters B, D, E and F individually showed significant bias in three traits. Though some of the raters displayed certain degree of differential severity/leniency in different traits of the rating scale, the infit and outfit mean squares for the raters were all within the acceptable range, indicating satisfactory internal consistency.

The bias/interaction can help to identify individual raters who exhibit misfit from expected ratings (Myford & Wolfe, 2004). Raters A, B and C were found to assign lower-than-expected ratings (from -0.13 logits to -0.21 logits) for the trait Speech Act. Raters A and B assigned higher-than-expected ratings for traits Expressions and Appropriateness (from 0.09 logits to 0.12 logits). Raters normally tend to assign higher ratings on easy trait and lower ratings to difficult trait. Considering that Speech Act (-0.38 logits) was the easiest trait and Expressions and Appropriateness (0.15 logits) were the most difficult traits, we expected raters to assign higher ratings on Speech Act and lower ratings on Expressions and Appropriateness. However, this was not true for Raters A, B and C. We found Raters A, B and C tended to assign lower ratings than would have been expected on the trait of Speech Act and Raters A, B tended to assign higher ratings than would have been expected on the traits of Expressions and Appropriateness. This evidence would suggest that Raters A, B, and C were to certain degree exhibiting halo effect.

5. Discussion

5.1. Discussion on the first research question

Do raters differ in the levels of severity in their WDCT ratings?

Data analysis indicated that the raters differed significantly in their level of severity. The most severe rater was Rater A (0.82 logits), the least severe rater was Rater E (-1.03 logits). Of all the six raters, four (Raters A, B, C and D) had logit values larger than 0.00 and two (Raters E and F) smaller than 0.00 logit.

Some studies have found that NNS raters tended to be more severe than NS raters (Fayer & Krasinski, 1987; Santos, 1988), while other studies revealed that NNSs were more lenient in many aspects than NSs (Brown, 1995; Liu, 2007). This study showed no contrast differences between these two groups. However, it did show significant differences between the two NS raters, Raters A and B. Rater A was the most severe rater while Rater B was much more lenient one. Rater B explained:

In fact, I have a Chinese girlfriend and she introduces to me some knowledge of Chinese culture. Moreover, I have stayed in China for eight years. I think that does affect my rating, because it makes me better understand some responses which may seem improper or inappropriate in our country.

This was in accordance with the findings in Liu's (2006) study, which indicated that raters' familiarity with the examinee's native culture affected the raters' behavior, normally towards the lenient side.

Different from Rater B, Rater A denied the influence of the Chinese culture in his rating. "I don't think the knowledge of Chinese culture would affect my rating. I just rated on his level of English." He owed his severity to other elements:

I don't think I rated much too severe. Maybe it is because I have a different way of using of *the rating manual; I have my own definition and criterion of the four traits and expectation of the proper response.*

Analysis revealed that Speech Act was the easiest trait for the examinees to get a high score. For the four NNS teachers, they were unanimous in that they often gave a comparative high score as long as the examinees said “sorry”, expressed their apology directly and properly. However, the two NS teachers took it in a different way. Rater A remarked,

I think it is not easy to assign a high score on the ‘speech act’. ‘Sorry’ is a so widely used word that it cannot be held that its appearance is equal to an expression of apology. For example, in the sentence ‘I’m sorry, but could you move your car?’ This is a ‘true’ request but not an apology. What I really care is the way he expresses his apology. I mean it is the sincerity and realization patterns that matter.

It was also found that the raters had certain kind of self-expectation to an appropriate response. Such an expectation was the result of their individual views towards the rating manual, different experiences in the real life and disparate thinking processes during the rating. If the performance of a student was not good enough to match the expectation of the rater, the rater would tend to assign a low score. Rater E (most lenient) reported that, before rating, she did the test to get a whole picture of the test and formed her own expectation to each different situation. She attached more attention to Expressions, their grammatical knowledge, though it was emphasized during training that that ungrammaticality was not an issue for the purpose of the study.

5.2. Discussion on the second research question

Do raters effectively and consistently employ the rating scales in their WDCT ratings?

The existence of central tendency can be detected from the measurement of the examinees. At the group level, it was hypothesized that all examinees shared the same performance measure after accounting for measurement errors, thus a non-significant chi-square value suggested a group-level central tendency effect (Myford & Wolfe, 2004). The result (χ^2=2454.4, *df*=37, *p*<.01) suggested that there was not a group-level central tendency. The examinee separation index indicated that there were nearly 13 statistically distinct strata of examinee performance. And the high degree of examinee separation reliability of (.99) implied that the raters could reliably distinguish among the examinees who were well differentiated in terms of their levels of performance. All these indicators did not suggest a group-level central tendency.

Previous researchers have claimed that some raters would deliberately assign middle categories in order to over-pursue the intra-consistency, which would result in the central tendency (Wang & Bian, 2012). The six raters in this study performed well in this aspect. First, they all knew the purpose and the function of their ratings. The test was not meant to be a high-stakes one, which eliminated the possibility of being required to assume any responsibility due to rating inconsistency or errors. Therefore, the adoption of a “play-it-safe” strategy was never considered. Second, all of them finished the rating in their own convenient time, which reduced their rating fatigue defined by some researchers who found that raters’ performance deteriorated over time due to fatigue and they might show lower levels of accuracy as they became tired over the course of the scoring project (Wolfe, Moudler, & Myford, 2001). However, the raters also reported certain kind of tendency to use the central scores, just as Rater A admitted:

Yes, 2 and 3 are most frequently used. The middle scores are more common that the 1, 4, 5. I do so because it is not good enough to obtain a 5. However it may also make sense though it may not be understood clearly. Therefore, it is the middle of the correct answer.

This opinion was also echoed by the other five raters.

5.3. Discussion on the third research question

Do raters efficiently differentiate between traits, that is, do raters show any evidence of halo effect?

The halo effect may be detected by analyzing the traits of the rating scale. A non-significant chi-square value may indicate that the traits are not significantly different in terms of their difficulties. It may also suggest a pervasive trend toward halo in the ratings of all raters (Myford & Wolfe, 2004). The significant chi-square value ($\chi2$=297.1, df=3, p<.01) and the high degree of trait separation reliability of 0.99 both suggested that there was not a group-level halo effect.

As to the individual rater, it was found that different raters exhibited differential severity in different traits. This was consistent with the finding of Chalhoub-Develle(1995), Cahlhoub-Develle& Wigglesworth (2005) and Liu (2007). For example, for Expressions and Appropriateness, Rater B stated:

I did not cast much attention on the grammaticality. As long as the examinees can get themselves successfully understood, I would assign a higher score on these two traits.

Moreover, they sometimes attached emotional elements such as sincerity and formality to the Speech Act, which made them much more critical in assigning high scores on this trait. “I mean, it is the sincerity and realization patterns that matter.” maintained Rater A.

In this study, Raters C and D had infit and outfit mean squares significantly less than 1, indicating that they might exhibit halo effects. When their rating records were referred to, it was not hard to find that the scores of four traits of one situation were almost the same and the difference among traits was one point at most. Rater C complained:

I found it hard to distinguish 2, 3 and 4 points. It is much too subjective to assign these three points properly…As to the halo effect, I admitted that I was so anxious that I finished rating hastily, which would to some extent reduce the reliability of my rating.

Meanwhile, Rater D explicated:

My *halo effect may be due to two things. First, it may lie in the rating results of former few students because in the beginning, I found myself somewhat bewildered in distinguishing the four traits. But it became much better as I continued rating. It sounds just like learning while rating. Second, I admit that I would tend to label the examinee who performed poorly in his former responses as ‘the one with poor ability’, which may to some extent influence my rating of his other responses. Likewise, good impression would lead me to assign high scores.*

Moreover, situational factors, such as the rating environment, the raters’ physical and emotional status were too essential to be ignored, which was also acknowledged by Daly & Dickson-Markman(1982) and Wang (2007). During the interview with Rater C, she apologized:

I feel so sorry for you because the time when I did the rating, I had something else urgent to accomplish too. But I did not want to prolong the rating for so long, so I felt a little bit anxious in my heart and finished the rating hastily.

Raters A, B and F also recognized that a favorable environment and a quiet mood were indispensable for a successful rating, just as Rater F remarked, “Considering that rating is a tough task both physically and mentally, desirable environment is a necessity in cultivating good physical and emotional status.” Therefore, controlling the rating quality during the rating process is worth further exploration. Rater training to avoid raters’ halo effect seems very important. Myford and Wolfe (2003) recommended that researchers should train raters to be aware of the halo effect and the impact it could have on their ratings so that they could attempt to guard against this tendency.

5.4. Discussion on the fourth research question

Do raters exhibit bias in their WDCT ratings?

When the interaction between the rater severity and the examinee ability was investigated, 87 significant bias interactions were found from all the six raters. Rater A was more lenient than expected on the examinees with higher ability, but more severe than expected on the examinees with lower ability. On the contrary, Raters B, D and E were more severe than expected on the examinees with higher ability, but more lenient than expected on the examinees with lower ability. Rater C was more lenient than expected on the examinees with lower ability, while Rater F was generally lenient to all examinees. Rater C also exhibited halo effect and had 15 (17%) biases with the examinees. She was also one of the raters who were the most likely to exhibit central tendency. “My poor rating performance may also be due to my inexperience in teaching and rating for I have been teaching for only two years”, explained Rater C. She could be labeled as a novice rater. Huot’s study (1993) demonstrated that experienced raters rated more coherently than the novice raters, which was acknowledged by Wolfe and Feltovich (1994) and Wolfe and Ranney (1996). This study also echoed the findings made by Huot (1993). Lacking in teaching and rating experience resulted in Rater C’s comparatively poor rating performance.

Different from the findings that raters tended to be more biased towards examinees with high ability (Schaefer, 2008; Wang, 2010), this study suggested that raters were inclined to be more severe or lenient bias towards lower ability examinees rather than higher ability examinees. Moreover, this study echoed some of the major findings by Kondo-Brown (2002) in that every rater’s bias pattern was different and the highest percentage of significant biased rater-examinee interactions was found among examinees whose ability was extremely high or low. In this study, they were Examinee 1 (with 3 biases) and Examinee 9 (with 4 biases).

The interview provided clues to the rater-examinee bias. First, raters’ different criteria of a high level of English led to some bias. Some might focus on grammatical knowledge (Rater E), while others (Raters A and B) might attach more importance on the realization pattern. In addition, some (Raters C, D and F) owed it to the gap between the given responses and their expected ones.

Differential severity in different traits resulted in certain bias, too. Raters A and B exhibited leniency towards Appropriateness and Expressions while severity towards Speech Act. Rater C was more critical on Expressions while Rater D on the Speech Act and Appropriateness. Rater E showed lower expectation of the Expressions, while Rater F did not deliberately focus on any of the four traits.

Disparate factors considered apart from the rating manual also brought some bias. Sincerity was the factor Raters A and B cared the most, just as Rater B put, “You know, it is some kind of feeling. It just can’t touch you.” Rater A added, “What I really care is the sincerity.” Raters C and D assumed that each situation was of unique character which required different degrees of apology. As to Rater E, emotional factors got the priority, especially her maternity. “As a young mom, I just could not be much too critical. Moreover, I would try to understand the response from the perspective of the examinee.” Raters C and F emphasized the attitude of the response.

6. Conclusions and implications

Analysis in this study revealed that raters, with different nationalities and educational and professional backgrounds, showed significant differences in terms of their rating severity, with a general tendency towards being severe. The most severe rater was Rater A, an NS English teacher while the most lenient rater was Rater E, a Chinese ESL teacher. Though they could effectively and consistently employ rating scales in their ratings, Raters A, B, and C showed certain degrees of halo effect. They demonstrated less efficiency in differentiating different traits. Even though the six raters showed consistency in their ratings, most raters were also found to exhibit certain bias across both traits and examinees.

Though it has long been believed that rating differences survive rater training, the function of rater training is too significant to be denied. This study renders the following implications for rater training, rating quality control and language teaching.

First, the realization of rater training should be flexible. It could be conducted before rating, while rating and after rating. Before rating, detailed analysis and explanation of the rating scale should be given to all raters. This process also welcomes hot discussion with an aim to achieve a unanimous acknowledgement of the rating scale. It could be a solution to the problematic fact that each rater bears an individual expectation or criteria towards each question, failing to apply the rating scale strictly. In addition, raters could be provided with information on common rater errors and be cautioned not to commit them. While rating, peer assistant and supervisory control should be adopted. These two measures can help to point out the rating errors a rater is committing. Feedback from the outside could motivate one to adjust himself when necessary for a more qualified rating result. Rater training could also be implemented after rating because the finish of the rating task does not necessarily mean the end of cultivation of good rating expertise. Contrarily, the development of good rating expertise is a both time and practice-consuming process. This period of training could be carried out in a form of forum where each rater is free to express his or her uncertainty, doubt, new recognition and gains. Moreover, it could also provide a good opportunity for the communication between expert raters and the novice ones.

Second, the establishment of an internalized set of criteria should be encouraged. This is a decisive factor accounting for the intra-reliability. To achieve this goal, enough time and guidance should be granted to each rater. Moreover, considering that the results of training may not endure for long after a training session, it becomes much more necessary to emphasize the function of the internalization of the criteria.

Third, anchor descriptions could be provided to better explain the content and standard that each category stands for. For example, in this study, some raters find it hard to distinguish the differences between 2, 3 and 4 points. Also strong desires have been expressed to have some examples to help better discriminate those three points.

Fourth, the knowledge on the rater effects should be introduced to the raters, training them to be aware of the rater effects and motivating them to guard against the rater effects. Some strategies could be taken to reduce the rater effects. For example, to avoid the halo effect, we may have every rater rate each examinee, each rater rating a specific trait instead of all traits.

Rater effects are a perennial and ubiquitous phenomenon (Eckes, 2005), and may manifest themselves in a variety of ways. The existence of rater effects may threaten the validity of decisions that are made based on those ratings (Wolfe, 2004). Pragmatics testing is still on its initial stage, researching into the reliability and validity of different test facets is thus especially important.

Acknowledgements

This project has been funded by the National Social Science Foundation of China

(12BYY060) and supported by the MOE Project of Key Research Institute of Humanities and Social Sciences in Universities (11JJD740012).

References

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

Bachman, Lyle F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*, 238-257.

Barron, A. (2003). *Acquisition in interlanguage pragmatics : learning how to do things with words in a study abroad context*. Philadelphia, PA: J. Benjamins Pub. Co.

Billmyer, K., & Varghese, M. (2000). Investigating instrument-based pragmatic variability: Effects of enhancing discourse completion tests. *Applied Linguistics*, 21(4), 517-552.

Bonk, William J., & Ockey, Gary J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89-110.

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

Brown, J. D. (2001). Pragmatics tests: different purposes, different tests. In K. R. Rose & G. Kasper (Eds.), *Pragmatics in Language Teaching*. New York: Cambridge University Press.

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes, 24*(3), 383-391.

Chalhoub-Deville, Micheline. (1995). A Contextualized Approach to Describing Oral Language Proficiency. *Language Learning, 45*(2), 251-281.

Cohen, Andrew D. (2008). Teaching and assessing L2 pragmatics: What can we expect from learners? *Language Teaching, 41*(02), 213-235.

Daly, J. A., & Dickson-Markman, F. (1982). Contrast effects in evaluating essays. *Journal of Educational Measurement, 19*, 309-316.

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

Eckes, Thomas. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*, 155-185.

Elder, Catherine, Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly, 2*, 175-196.

Fayer, J. M. , & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning, 37*, 313-326.

Golato, Andrea. (2003). Studying compliment responses: a comparison of DCTs and recordings of naturally occurring talk. *Applied Linguistics, 24*(1), 90-121.

Gyagenda, Ismail S., & Engelhard, George, Jr. (1998). *Applying the Rasch Model To Explore Rater Influences on the Assessed Quality of Students' Writing Ability.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego.

Hsieh, Ching-Ni. (2011). *Rater Effects in ITA Testing: ESL Teachers' versus American Undergraduates' Judgments of Accentedness, Comprehensibility, and Oral Proficiency.* (PhD), Michigan State University, Michigan.

Hudson, Thom, Detmer, Emily, & Brown, James Dean. (1995). *Developing Prototypic Measures of Cross-cultural Pragmatics*. Honolulu: Second Language Teaching and Curriculum Center, University of Hawaii at Manoa.

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations*. Cresskill, NJ: Hampton Press, Inc.

Kasper, G. (1998). Interlanguage Pragmatics. In H. Byrnes (Ed.), *Learning Foreign and Second Languages: Perspectives in Research and Scholarship* (pp. 183-208). New York: The Modern Language Association of America.

Kondo-Brown, Kimi. (2002). A FACETS Analysis of Rater Bias in Measuring Japanese Second language writing performance. *Language Testing, 19*(1), 3-31.

Kumar, DSP Dev. (2005). Performance appraisal: The importance of rater training. *Journal of the Kuala Lumpur Royal Malaysia Police College, 4*, 1-17.

Linacre, J.M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2012). *A User's Guide to FACETS: Rasch-model Computer Program*. Chicago: MESA Press.

Liu, Jianda. (2006). *Measuring Interlanguage Pragmatic Knowledge of EFL Learners*. Frankfurt am Main Peter Lang.

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

Lumley, T., & Brown, A. (2005). Research methods in language testing. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning* (pp. 833-855). Mahwah, NJ: Lawrence Erlbaum.

Lumley, Tom, & McNamara, T. F. (1995). Rater Characteristics and Rater Bias: Implications for Training. *Language Testing, 12*(1), 54-71.

Myford, Carol M., & Wolfe, Edward W. (2003). Detecting and measuring rater effects using Many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.

Myford, Carol M., & Wolfe, Edward W. (2004). Understanding Rasch measurement: detecting and measuring rater effects using Many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.

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

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the Ratings: Assessing the Psychometric Quality of Rating Data. *Psychological Bulletin,* 88(2), 413-428.

Santos, T. (1988). Professors'reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly, 22*, 69-90.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493.

Scullen, S.E., Mount, M.K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*, 956-970.

Taguchi, N.(Ed.). (2009). *Pragmatic competence*. Berlin ; New York: Mouton de Gruyter.

Taguchi, N. (2011). Rater variation in the assessment of speech acts. *Pragmatics* 21(3), 453-471.

Walters, F. S. (2007). A conversation-analytic hermeneutic rating protocol to assess L2 oral pragmatic competence. *Language Testing*, 24(2), 155-183.

Wang, B., & Bian, R. (2012). The mechanism of conservativeness in subjective performance rating. *Psychological Exploration, 5*, 429-438.

Wang, H. Z. (2007). Rater perceptions of factor that affect the rating of TEM-4 oral test. *CELEA Journal, 30*(2), 9-15.

Wang, Z. F. (2010). *A study of rater bias in rating the NMET writing performance.* (MA), Shanxi University.

Wiseman, Cynthia S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing, 17*(3), 150-173.

Wolfe, E. W., & Ranney, M. (1996). Expertise in essay scoring. In D. C. Edelson & E. A. Domeshek (Eds.), *Proceedings of ICLS 96* (pp. 545-550). Charlottersville, VA: Association for the Advancement of Computing in Education.

Wolfe, E.W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 1*, 35-51.

Wolfe, E.W., Moudler, B. M., & Myford, C.M. (2001). Detecting differential rater functioning over time (DRIFT) using a Rasch multi-facted rating scale model. *Journal of Applied Measurement, 2*, 256-280.

Wolfe, Edward W., & Chiu, Chris W. T. (1997). *Detecting Rater Effects with a Multi-Faceted Rating Scale Model.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

Wolfe, E. W., & Feltovich, B. (1994, 0401). *Learning to Rate Essays: A Study of Scorer Cognition.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

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

Youn, S. J. (2007). Rater bias in assessing the pragmatics of KFL learners using Facets analysis. *Second Language Studies*, 26(1), 85-163.