Journal of Language Testing & Assessment (2018) Vol. 1: 1-13
Clausius Scientific Press, Canada

To What Extent Do Non-teacher Raters Differ from Teacher Raters on Assessing Story-retelling

Lu Weilie

Guangdong University of Foreign Studies, South China Business College, Guangzhou, China
Email: 891285101@qq.com

Keywords: Non-teacher Raters, Teacher Raters, story-retelling, consistency, severity, written comments

Abstract: The present study aims to explore to what extent non-teacher raters differ from teachers' raters on assessing story-retelling in China's National Matriculation English Test, Guangdong version. Facets analysis suggests that the two rater groups are comparable in terms of internal consistency and severity. Results from raters' written comments show that both rater groups followed a similar pattern in focusing on different criteria categories when making assessments. Difference occurs only in how they commented: teacher raters' comments tended to be more specific while non-teacher raters' comments were more general. Based on the findings, we conclude that in rating high stakes tests like Story-retelling in NMET GD, non-teachers (college/graduate students) are qualified to be raters.

1. Introduction

Since 2011, NMET GD(National Matriculation English Test, Guangdong Version) has incorporated a new performance test, a computer-based English listening and speaking test (CELST). It is an integrated test of candidates' listening and speaking abilities. It consists of three parts: Part One is reading-aloud, which mainly tests candidates' pronunciation; Part Two is a role play, which tests the students' communication efficiency and grammar and vocabulary use (Zeng2009:236), and Part Three is story-retelling, which tests candidates' integrated listening-speaking ability. In story-retelling, candidates first listen to a story twice, while their listening, five key words in Chinese and a one-sentence gist of the story are presented on the computer screen to help them better understand the story and at the same time candidates are encouraged to take notes. After the second listening, candidates have to retell the story within one minute.

The inclusion of such a performance test in NMET GD is considered to be an advance because it can bring positive washback effects to classroom teaching and learning. However, the involvement of human raters in giving scores to candidates poses some threats to the reliability and validity of the test. What's more, it will also put greater burden on teachers since they have to spend much more time in rating students' performance on this test and give feedback to them in their mid-term exam, final exam or even in their daily practice.

As to the question "who are qualified to be raters", results differ. According to Brown (1995), it

is generally the case that experienced teachers only are permitted to train as assessors for language tests (Brown, 1995). In the study conducted by Shohamy at el (1992), they acknowledged that teacher raters and lay raters may rate differently in that teachers may be influenced by instructional goals and may emphasize particular components of the written sample while lay raters may rate the written sample based on their personal views of the written language. However, it was found that lay raters can rate as reliably as professional raters. The suggestion they made at last is that 'decision makers, in selecting raters, should be less concerned about their background, since that variable seems not to increase reliability.' Schoonen at el (1997) differentiated free assignments from restrictive assignments. Free writing assignments elicit very diverse essays, which complicate the rating task and thus demand a great deal of expertise when rating, while restrictive assignments elicit a limited range of respone, which simplifies the rating task and thus and easily be carried out by non-teachers non-experts. Their results showed that in rating restrictive tasks lay raters were as reliable as expert raters.

Story-retelling, whose answer or response is controlled by the information or language material provided, also belongs to restrictive productive task(Li Xiaoju,1997:146). When rating this task, to what extent do non-teacher raters differ from teacher raters? This is what the present writer intends to explore.

2. Research Questions

The use of speaking in assessment necessitates the use of raters to score test-takers' performance, based on their own judgments, which makes the reliability issue more complicated because human judgments involve subjective interpretation on the part of raters and may thus lead to disagreement (McNamara, 1996). Reliability is a prerequisite to validity in performance assessment in the sense that the test must provide consistent, replicable information about candidates' language performance (Clark, 1975, cited in He Manzu, 2006). According to Jones (1979), reliability in a performance test depends on two variables: 1) the simulation of the test tasks; 2) the consistency of ratings. According to McNamara (1996), consistency refers to the extent of the random error associated with raters' ratings. When inconsistency occurs, 'it thus becomes difficult to model the raters' characteristics, and thus to build in some compensation for them….and such raters, once identified, may need to be retrained or, failing this, excluded from the rating process'. And Wiseman (1949) regarded self-consistency as the most crucial quality in a rater, without which, 'no orderly process of measurement can be conducted' (Lumley & NcNamara, 1995)

Rater severity is one of the most prevalent rater effects in performance assessment (Ching-Ni. Hsieh, 2011). It is the effect that Cronbach (1990) referred to as the most serious error that a rater can introduce into a rating procedure. The effect occurs when raters provide ratings that are consistently either too harsh or too lenient as compared to other raters (Lumley & Mcnamara, 1995; McNamara, 1996). It is a common rater effect among raters which will persist even after training (Brown, 1995). Since such disagreements tend to depress the correlation among raters and thereby lower score reliability, it is necessary to find out whether raters differ in their overall severity, once detected, then compensation for rater characteristics needs to be built into the rating process (Manamara, 1996).

Given the importance of self-consistency and severity in rating performance test, researchers often take them into consideration when commenting on raters' rating quality (Kim,2009, Ying Zhang and Catherine Elder's, 2010). In those studies, researchers compared raters' consistency and severity focusing on scores they produced. However, such scores are the products, from which we do not understand what process raters went through during their rating. As Connor-Linton (1995) pointed out, 'if we do not know what raters are doing…then we do not know what their ratings

mean' (P195). Researchers, therefore, have begun to call for more in-depth investigation into raters' decision-making to explore their rationales when judging the target performance and the thought processes which go through their mind, which might help to account for the idiosyncrasies detected by the statistical modeling of the final scores. (Douglas 1994; Connor-Linton 1995; Cumming 1997, cited in Zhang jie, 2009). And raters' written comment is one method commonly used by researchers (Ling Shi, 2001;Kim, 2009; Yingzhang and Elder's, 2010). It is considered to be a helpful and essential way to learn about the process of raters' interpreting and applying the rating scale (Binhong Wang,2010).

Based on the previous studies and being aware of the possible existence of difference between the assessments made by the two rater groups, in this research, I aim to find answers for the following research questions:

1)To what extent do non-teacher raters and teachers differ in consistency?

2) To what extent do non-teacher raters and teacher raters differ in severity?

3) To what extent do non-teacher raters and teacher raters differ in their written comments?

3. Method

3.1 Raters

Two groups of raters were invited to take part in this research. The first group consisted of 10 teachers. They all had extensive experience in rating performance tests (e.g. Business English Certificate, NMET GD, placement tests, classroom-based achievement tests). There are 3 males and 7 females. Their ages ranged from 30 to 39. They all had academic backgrounds in language education or linguistics and experience teaching ESL for no less than 3 years. The second group included 10 graduate students majoring in applied linguistics, foreign language teaching or language evaluation. All of them passed TEM 8. There are also 3 males and 7 females. Their ages were more homogeneous, ranging from 22 to 24. They reported having neither rating experience nor teaching experience.

3.2 Material

The story-retelling samples used for the present study were produced by examinees when they were taking a simulation test in preparation for the GD NMET just one month before the operational speaking administration. Altogether 30 samples were collected including low, middle and high level performance.

Raters assessed candidates's performances from two different aspects. The first aspect focused on 10 information points(IP), 1.5 marks for each, with four possible scales for each IP, that is 0,0.5,1 and 1.5. The second aspect is Holistic Impression. Raters will consider Content, Language, Fluency and Pronunciation. Holistic mark ranges from 0 to 9. The original total mark for the story-retelling is 24. However, when added to the final NMET result, it becomes 6.

In this study, each rater altogether rated 30 candidates' retelling samples. When rating the first 24 candidates' samples, raters awarded their scores by circling the correspondent mark on the rating sheet. And from the 25th candidate to the 30th candidate, raters were required to not only award their scores but also give their reasons for that score on a written comment sheet .

3.3 Data collection

Data collected for the study includes a questionnaire about raters' demographical information, the scores each rater gave to the 30 speech samples and the written comments justifying their

awarding a particular score. Before analysis, all the ratings that the two rater groups had provided (24 marks in total) were divided by four, to make the total score 6, which was the way they would be treated in the operational NMET GD.

In order to answer the first and second research questions related to self-consistency and severity, data were subjected to analysis using many-faceted Rasch measurement (MFRM). The computer programme, FACETS 3.58, was used for this analysis.

In order to answer the third research question 'To what extent do the two types of raters differ in their written comments?', the raters' written reasons were analyzed based on the methodology used by Kim (2009) and Wang (2008).

4. Results

4.1 All Facets summary

Figure 1 is a graphical summary for the measures of all facets included in the analysis. The leftmost column stands for the common scale in the unit of logit, against which all the measures in the following facets are calibrated. The second column compares the 20 raters in terms of the level of severity/leniency. The higher the raters are, the more severe they are when rating the story-retelling. From this figure, we can know that the 20 raters were about 3.3 logits apart in their relative severity, with sr6 being the most severe rater, and sr7 and tr5 the most lenient. The third column displays the distribution of estimates for examinees' retelling ability with those who are more competent listed at the top while the less competent ones at the bottom. Of all the 30 candidates, 24 of them are above 0 logit and only 6 below 0 logit, showing that most of the candidates are quite competent in the task or that the retelling task is easy for the candidates. From the second and third column, we know that the examinee facet has much wider span on the logit scale than rater facet, suggesting that the most significant part in score variance lies in the examinees' ability rather than the raters.

4.2 To what extent do non-teacher raters and teacher raters differ in consistency?

Infit Mnsq indicates to what extent the raters could rate self-consistently, as the model expects. There is no hard-and-fast rule for interpreting fit statistics. Considering the small number of ratings each rater awarded, a lenient criteria (0.6--1.4) is adopted for detecting raters with rating problems. (Zhang Jie&He Lianzhen, 2008). Infit mean square values greater than 1.4 indicate significant misfit, or a high degree of inconsistency in the ratings. In contrast, infit mean square values less than 0.6 indicate overfit, or a lack of variability in their scoring.

The most relevant figures to this study is the 6th column in Table 1, which shows the fit statistics for each rater.

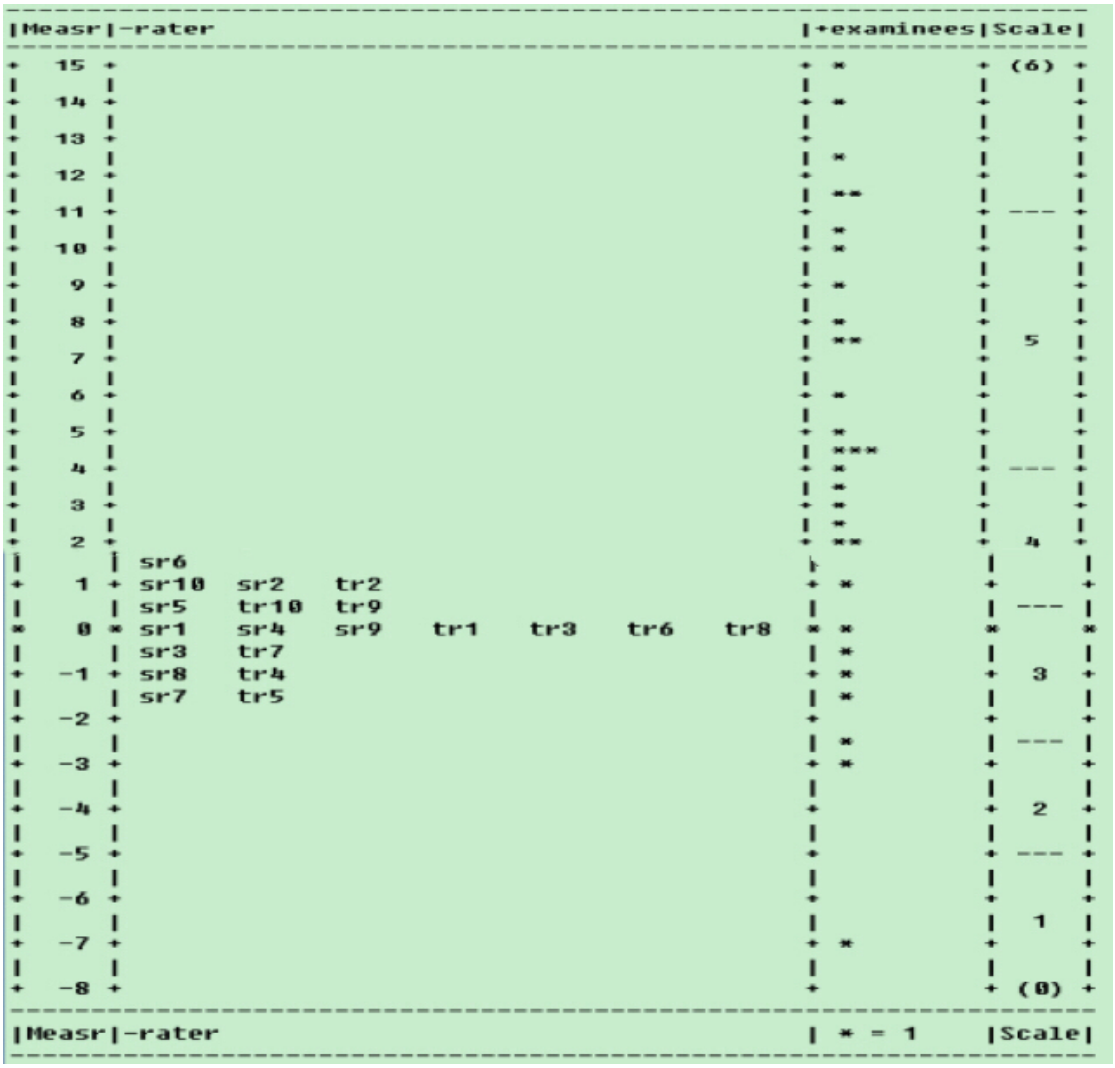

Figure 1 All facets summary

Notes: tr stands for teacher rater; sr stands for non-teacher raters

From this column, we can see difference exist between the two rater groups. In the non-teacher rater group, one rater is spotted as being misfit: the fit value for NTR1 is 1.41, a bit larger than the upper limit, indicating that there is a inconsistency in the ratings. In the teacher-rater group, TR7, whose fit value is .60, falls on the borderline of the lower limit, implying that there might be a lack of variability in his/her scoring. Except for the above minor difference, all the other 9 raters in both the non-teacher rater group and the teacher rater group fall within the acceptable limits. Table 2 is the summary of this result.

Table 1 Rater Measurement Report

Raters	Obsvd Average	Fair-M Average	Measure	Model S.E	Infit MnSq	ZStd
NTR1	4.3	4.63	-.10	.39	1.41	1.4
NTR2	4.0	4.31	1.05	.37	.67	-1.2
NTR3	4.3	4.66	-.26	.39	.69	-1.1
NTR4	4.2	4.55	.19	.38	.66	-1.3
NTR5	4.2	4.51	.34	.38	.66	-1.3
NTR6	1.0	4.24	1.32	.37	.65	-1.3
NTR7	4.6	4.89	-1.56	.42	1.05	.2
NTR8	4.6	4.85	-1.22	.41	.95	.0
NTR9	4.3	4.63	-.10	.39	.79	-.7
NTR10	4.0	4.31	1.05	.37	.95	-.1
TR1	4.2	4.55	.19	.38	.80	-.7
TR2	4.1	4.35	.91	.38	.92	-.2
TR3	4.2	4.55	.19	.38	..66	-.13
TR4	4.6	4.85	-1.22	.41	.99	.0
TR5	4.6	4.89	-1.56	.42	1.05	.2
TR6	4.3	4.59	.04	.39	.71	-1.0
TR7	1.3	4.66	-.26	.39	.60	-1.5
TR8	4.3	4.59	.04	.39	.71	-1.0
TR9	4.2	4.47	.48	.38	1.14	.5
TR10	4.2	4.47	.48	.38	1.34	1.5
Mean	4.3	4.58	.00	.39	.87	-.5
S.D.	.2	.19	.84	.02	.24	.9
RMSE(model)=39 Adj. S.D.=.75 Separation= 1.91 Separation (not inter-rater) Reliability=.79 Fixed (all same) chi-square= 84.6 df =19 Significance (Probability)= .00						

Table 2 Comparison of internal consistency between groups

Rater Type	Number	Within acceptable limits	On/beyond borderline of acceptable limits
TR	10	9	1
NTR	10	9	1

Besides fit statistics, ZStd (Z standardized) value can help to test whether raters are self-consistent or not. Generally speaking, fit statistics and ZStd value are combined to judge raters' consisency. Z value larger than 2 shows significant misfit and less than -2 shows significant overfit(Li Qinghua&Kong Wen, 2010). From the last column ZStd in Table, we can see that none of the raters, no matter raters from non-teacher rater group or raters from teacher rater group, show ZStd value larger than 2 or less than -2, indicating that raters in both of the rater type can be self-consistent while rating the story-retelling.

In summary, although one rater in the teacher group and one rater in the non-teacher group are suspected as being overfit or misfit, the Z value implies that the two groups rarely differed in terms internal consistency.

4.3 To what extent do non-teacher raters and teacher raters differ in severity?

Generally speaking, a separation ratio of more than 2 and reliability index larger than 0.9 could well indicate statistically significant difference among all the elements, which in the rater facet means that there is significant and consistent difference in rater severity among the raters (Zhang Jie & He Lianzhen, 2008). From the bottom of Figure 2, we know that the separation index and reliability are 1.91 (less than 2) and 0.79 (less than 0.9) respectively, which implies that on the whole there is no significant and consistent difference in rater severity among all the raters. To put it another way, teacher raters do not differ significantly from non-teacher raters in severity.

To confirm this result, we further examine the severity measure. For each rater, the measure column refers to their level of severity across all the examinees they rated. In Table 1 the 4th column *measure* shows the logit values for each rater. Logit values for rater severity ranged from -1.56 to 1.32.

The rater logit severity calibrations were compared for rater groups in order to determine whether one type of rater rated significantly more harshly than the other. The *t*-test was carried out on the teacher rater vs. the non-teacher raters.

Table 3 Severity comparison between groups

	Mean logit value	Standard deviation	Standard error
NTRs (n=10)	.071	.946	.299
TRs (n=10)	-.071	.767	.243
t-value =.401 df = 9, ns.			

From Table 3, we know that the mean logit value and the standard deviation for the non-teacher raters are .71 and .946 respectively, while the correspondent values for the teacher raters are -.071 and .767. Although non-teacher raters as a group were found to be more severe than teacher groups (.071 vs. -.071), the difference was minimal and non-significant, suggesting that the two rater groups are comparable in their severity when rating the retelling task.

To clonclude, there is minor difference between non-teacher rater group and teacher raters group: the non-teacher raters are found to be a bit severe than their counterparts. However, a t-test comparing the harshness between the two rater types showed that the difference was minimal and non-significant. Besides, the low separation value (1.91) and the low reliability value (.79) implied the comparable severity between the two rater groups.

4.4 To what extent do non-teacher raters and teacher raters differ in their written comments?

When assessing the 1st to the 24th candidates, raters just awarded a score. For the 25th to the 30th candidates, besides awarding scores, raters also provided their written comments justifying the marks they had awarded. For each of these candidates, raters provided their written comments for the 10 IPs as well as the Holistic impression. After going over all the raters' comments, it was found that raters commented quite differently in these two aspects. They supplied just one phrase like "*information correct*", "*information incorrect*" or at most one sentence justifying their awarding a particular score. When they commented the Holistic Impression, they provided more comments explaining why they awarded that score. Therefore, it was decided that the qualitative results will be compared from these two different aspects, first, comparison will be made on comments on IPs, and then comparison will be made on comments on Holistic Impression.

4.4.1 Results of written comments on IPs

For each IP, a maximum mark is 1.5. And the possible scores a rater can award are 0, 0.5, 1 and

1.5. A mark of 1.5 means that such a piece of information is totally correct while a mark of 0 means the opposite. So when commenting on 0-mark or 1.5-mark IPs raters used nearly the same words. Therefore, the comparison of our focus was on the 1-or-0.5-mark IPs.

Of all the written comments on all the IPs, non-teacher raters and teacher raters provided the similar number of comments on the 0.5-or-1 mark IPs: 130 vs. 140. However, after careful examination, it was found that the way they commented differed greatly. Teacher raters tended to be specific while non-teacher raters only provided some general comments. By '*general*', we mean that the comments only touch upon a certain rating criteria, with no further words to pinpoint why the candidate did or didn't do well in this aspect. If the rater can pinpoint which aspect the candidate did or didn't do well in, we classify their comments into 'specific'. Table 4 provides examples to show the way we distinguish '*general comments*' from '*specific comments*'.

Table 4 Examples of General comments vs. Specific comments

General comments	**Specific comments**
Information missed	*only retell one of the two pieces of information, do sth about it was left*
Information incorrect	*only mentioned got angry, but without decided to do sth*
Not loyal to the content	*Information basically correct, but it didn't point out the location 'in his shop'*
Basically correct	*Information basically correct, but it didn't mention the time 'the next morning'*

Table 5 shows the results of the frequency of 1-or-0.5-mark IPs, the frequency of specific comments provided and the result of a Chi-square test.

Table 5 Difference of specific comments between groups

	Frequency of 1-or-0.5-mark IPs	Specific comments	X^2	Sig Level
NTRs(n=10)	130	35	12.63	.000
TRs (n=10)	140	85		

Although both rater groups provide similar number of comments to 1-or-0.5-mark IPs, teacher raters provided far more specific comments than the non-teacher raters (85 vs. 35). And the Chi-square test showed that the difference was significant (X^2 12.63, Sig Level .000). Such a result may suggest that teacher raters are more evidence-based when giving a particular score. They may be clearer why they awarded a particular score. On the contrary, providing fewer specific reasons indicates that non-teacher raters are not as clear as their counterparts. They may have a rough idea, but difficult for them to tell the exact reason.

4.4.2 Results of written comments on Holistic Impression

Before comparing the comments between the two rater groups, the first thing is to decide how to code their comments. Following Kim's approach (Kim, 2009) and Wang's approach (Wang, 2007), the written comments were analyzed based on rating criteria. Four general evaluation criteria were first abstracted: Content, Language, Fluency and Pronunciation. However, after comments from both groups were reviewed again and again, sub-categories emerged from the categories of Content and Language: under the category of Content, we further specified Completeness (when raters comment on whether the retelling is complete or not), Logic (when raters mention the logic of the retelling), Coherence (when raters comment on the this aspect on the retelling) and Loyalty (when raters stated whether the retelling is loyal to the original one or not); under the category of Language, Language-general (when raters commented on the retelling on language in a general way

like ‘language is good’), Grammar, and Appropriateness were further specified. Table 6 presents the four main categories and seven sub-categories we used to code the raters’ written comments as well as the respective examples from raters’ comments.

Table 6 Categories to code the written comments

Category		**Exemplary comments**
Content	Completeness	‘*The retelling is complete in content’* ‘*Some information is missing*’
	Loyalty	‘*The retelling is not loyal to the original’* ‘*The retelling differs from the original version’*
	Logic	‘*The logic of the retelling is fairly clear*’
	Coherence	‘*Good coherence*’
Language	Language-general	‘*Language is good*’
	Grammar	‘*There are a lot of grammar mistakes. Tense misused.*’
	Appropriateness	‘*The expression is appropriate*’
Fluency		‘*The candidate is fluent in retelling the story’*
Pronunciation		‘*Pronunciation is quite good*’

Based on the above nine categories, the present writer coded all the written comments. Then a second coder (a graduate student whose research interest is also language testing) was invited to recode 1/4 of the comments. Our results reached approximately 95% agreement, which assures the practicality of such a coding method and the reliability of the present writer’s coding.

The comments were compared first across the two rater groups through a frequency analysis. According to the rating criteria used for the NMET GD, we have abstracted four main categories. First we want to know whether the two rater groups differed in frequency of their written comments across the four main categories. Table 7 reports the raw frequency counts for the four main categories.

Table 7 Frequency of main categories in comments between groups

	Content	Language	Fluency	Pronunciation	Total
NTRs	100	55	51	45	251
TRs	102	60	56	50	258

From the above table, we know that of all the four main categories, when assessing candidates performance of story-retelling with the Holistic approach, teacher raters focused on Content most frequently (100), then came to Language (55), Fluency (51) and Pronunciation (45). The same pattern occurs when examining non-teacher raters’ comments, whose frequency across the four main categories are 102, 60, 50 and 60 respectively. Such similar distribution of frequency on the four main categories pre-specified in the rating criteria seems to suggest that when awarding scores, both rater groups tend to focus on the same features of the retelling performance.

However, after closer examination, difference occurs as to distribution of their comments on sub-categories.

Table 8 Frequency of comments on sub-categories

	Content				Language		
Sub-categories:	completeness	loyalty	logic	coherence	language-general	grammar	appropriateness
NTR	41	36	14	9	10	28	17
TR	37	33	19	13	11	30	19

From Table 8, we know that the non-teacher group provided more comments than the teacher

group for two sub-categories: completeness (41vs.37), loyalty (36 vs. 33), and that the teacher group provided more comments than the non-teacher group for another five sub-categories: logic (19 vs. 14), coherence (13 vs. 9), grammar (30 vs. 25), appropriateness (19 vs. 10) and language-general (11 vs. 10), which may suggest that when awarding scores to these two main categories, non-teachers will focus more on completeness, loyalty and language-general while less on logic, coherence, grammar and appropriateness than teacher raters, which can be seen more clearly from Figure 2.

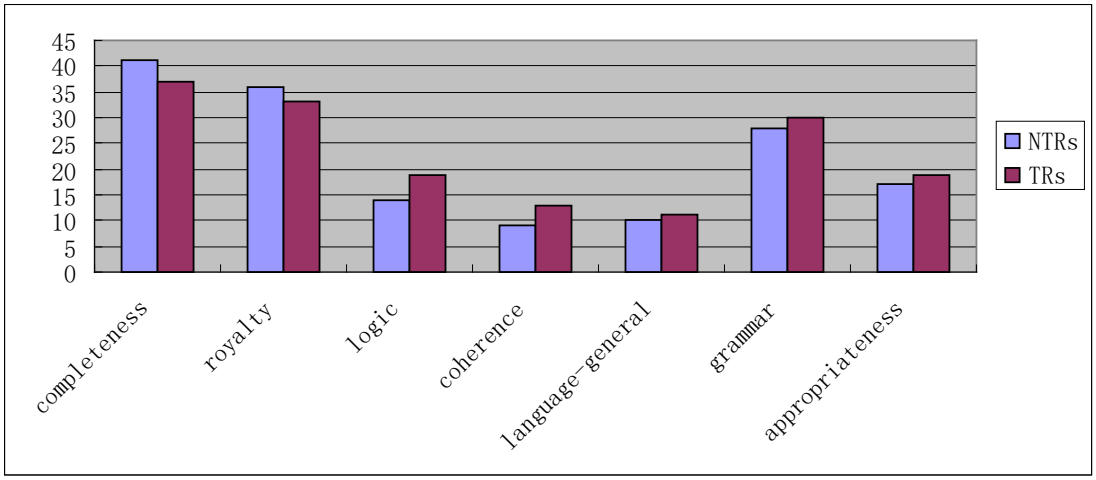

Figure 2 Frequency of comments on sub-categories

According to Table 8 and Figure 2, under the main category of Content, the non-teacher group was found to draw most frequently on completeness (41% of all comments on Content), second comes royalty (36%), then logic (14%) and coherence (9%). Teacher rater group showed the same pattern, who comments most frequently on completeness (36% of all comments on Content), second on loyalty (32%) and then comes logic (19%) and coherence (13%). These trends indicate that the two rater groups shared common ideas about the way in which the candidates' performance should be assessed on the category of Content. When they award scores to the candidates on the category of Content, completeness of the story and whether the retelling is loyal to the original story or not are two features that most draw raters' attention. Then come logic and coherence.

Language is the second most frequently commented on of all the four main categories. Under this category, there are three sub-categories: Language-general, Grammar and Appropriateness. According to Table 8 and Figure 2, the sub-category that most drew non-teacher raters' attention is Grammar (50.9% of all comments on Language), the second is Appropriateness (30.9%) and at last comes Language-general (18.2%). Again, teacher raters showed a similar pattern when commenting on these sub-categories. What most drew the teacher raters' attention is Grammar (50% of all comments on Language), the second is Appropriateness (32%) at last comes Language-general (18%). Such a similar pattern may suggest that the two rater groups agree as to what to focus on when making assessment on the categories on Language. For both the non-teacher raters and the teacher raters, Grammar is the most important, Appropriateness is the second most important and at last what they will consider is Language-general. The fact that both the student raters and teacher raters tended to focus more on Grammar may have something to do with their experience in the context of Chinese education. Teachers, who prepared for different kinds English tests when they themselves were students, are now helping students pass examinations. And in China's teaching context, grammar is still the focus of teaching and testing, especially in middle schools. Students, with an aim to achieve higher scores in order to enter their ideal higher school or institute, have to learn grammar by heart. Therefore both teacher raters and student raters may be most sensitive to

this aspect when assessing candidates' retelling performance.

For most of the comments, the two rater groups showed a similar pattern. However, when looked at more closely, again, it was found that teacher raters tended to provide more specific and elaborate comments, while non-teacher raters tend to comment in a more general way. For example, when evaluating Candidate 30 on the category of Content, non-teacher raters only provided some general terms like 'content basically loyal to the original', 'miss some information', 'content complete' etc. Different from their counterparts, the teacher raters' comments were more specific and elaborate, including 'it is good at the beginning, however, when it comes to the later part of the retelling, some information was missed' 'basically speaking, the candidate retell the original story, however, the plot of the story was a big confusing' 'the original story didn't mention tailor B's going into tailor A's shop, this part is made up by the candidate' etc.

When commenting on Pronunciation the same pattern appeared: the teacher rater group's comments tended to be more specific and elaborate. In their comments, they provided explicit pinpointing of pronunciation errors like 'on the whole, pronunciation is good, but the pronunciation of some particular words is not satisfactory, such as *attract, tailor*.' 'the mispronunciation of the word tailor will more or less affect listeners' understanding of the story', and they could also point out the reason for their poor pronunciation "influenced by dialect, there exist heavy dialect accent". Besides, when commenting on excellent pronunciation, they will also express their feelings for its 'pronunciation is really wonderful. It is fairly comfortable to listen to it'. When it comes to look closely at non-teacher raters' comments, a different picture appeared. Non-teacher raters were more general. They tended to focus on the overall quality of students' pronunciation performance. For example, their comments included "pronunciation is natural" 'pronunciation is relatively good' 'pronunciation is rather bad/poor'. Teacher raters' specific and elaborate comments on pronunciation might imply that the teacher raters tended to be more sensitive in this category. It can also be interpreted to suggest that teacher raters, all of whom have teaching experience of more than five years, who have heard and judged different students oral English, are more skillful in evaluating this aspect. While for the non-teacher raters, who have no teaching experience, they might distinguish good pronunciation from poor pronunciation, however, they are not as skillful as the teacher raters to pinpoint the problem.

5. Conclusion

Enlightened by the research done by Brown (1995), Schoonen et al (1997), Wang (2008), Kim (2009) and Zhang & Elderr (2010) and other researchers, the present study has found out:

(1) Although there is minor difference, result showed that high reliability in internal consistency can be achieved by both the non-teacher group and the teacher group. There was no significant difference between the two groups as long as internal consistency was concerned.

(2) In terms of harshness, both groups produced comparable ratings for the 30 candidates. Although non-teacher raters' score was a little bit higher, the difference is not significant.

(3) Both the non-teacher rater group and the teacher rater group produced tended to justify their scores in the similar way. They provided similar number of comments (251 vs. 258). They both drew most frequently on Content, then on Language, Fluency and Pronunciation. Difference occurs only in that teacher raters' comments were more specific, which couldn't be used as evidence to claimed non-teacher raters as unqualified raters.

The preceding findings demonstrated the comparability of non-teacher raters and teacher raters in terms internal consistency, severity and justifications, when rating story retelling in the NMET GD. Several implications can be drawn from this research.

First, the practical implication for this finding is that the tradition to have students' performances

rated by panels of expert raters is not necessary. In rating restrictive tasks, non-teacher raters, who are less costly and are more willing, can be recruited because their ratings are as comparable as those teacher raters. Just as Shohamy et al (1992) proposed, 'decision makers should have the option of selecting an economical group of raters …..'. On the one hand, it can lessen the burdern of teachers and make the test less expensive, on the other hand, rating candidates' performance is a useful experience for those students who plan to be teachers.

Second, there is a practical implication for teachers who prepare students for the NMET GD. As the qualitative results suggested, both groups of raters drew most frequently on Content, Language. And when awarding scores on the aspect of Content, both groups of raters focus most on Completeness and Loyalty. This can give some light to teachers in what aspects they should focus on when guiding their students in preparation for such a test. Since Content is the most important factor for scores, teachers should instruct students in how to take as many notes as possible while their listening. Since Completeness and Loyalty are the most and second most important features raters will pay attention to, teachers should warn their students of not trying to be creative in the story-retelling. Instead, teachers should encourage students to listen to the original story attentively and retell as closely in meaning as possible to the original version, never try to make up their own story. Through repetitive practice, their skill of taking notes may be improved, leading to covering more IPs in the story retelling and being loyal at the same time.

Third, in tests which feedbacks for students are required, teacher raters seem to be more suitable because they can provide more specific reasons, which may be useful for students' future study.

Acknowledgement

The present study is sponsored by the project *Guangdong Characteristic Key Subject English Construction Project* (GDTX170109).

References

[1] Anne Brown (1995) The effect of rater variables in the development of an occupation-specific language performance test. Language Testing, 12 (1): 1—15.

[2] Arthur Hughes and Chryssoula Lascaratou (1982) Competing criteria for error Binhong Wang (2010) On Rater Agreement and Rater Training, English Language Teaching Vol. 3, No. 1.

[3] Ching-Ni Hsieh (2011) Rater effects in ITA testing: ESL teacherrs' vesus American undergraduates' judgements of accentedness, comprehensibility and oral proficiency, Spaan Fellow Working Papers in Second or Foreign Language Assessment, Volume 9: 47—74.

[4] Connor-Linton, J. (1995). Looking behind the curtain: what do L2 composition ratings really mean? TESOL Quarterly 29 (4): 762-765

[5] Cronbach, L.J. (1990) Essentials of Psychological Testing (5th ed.). New York: Harper and Row. Cumming. (1990) Expertise in evaluating second language compositions, Language Testing, 7, pp31-51.

[6] Elana Shohamy, Claire M. Gordon and Robert Kraemer (1992) The effect of Raters' Background and Training on the Reliability of Direct Writing Tests, The Modern Language Journal, Vol. 76, No. 1 (Spring, 1992), pp. 27—33.

[7] Hadden, B.L. (1991) Teacher and nonteacher perceptions of second-language communication. Language Learning, 41, 1—24.

[8] He Manzu. (2006) A FACETS Analysis of Rater Bias in Measuring Chinese Students' English Writing—A Comparative Study of Holistic and Analytic Scoring Methods, Unpublished MA thesis, Guangdong University of Foreign Studies, China

[9] Li Xiaoju. (1997) The Science and Art of Language Testing Hunan Education Publishing House Lumley & NcNamara. (1995) Rater characteristics and rater bias: implications for training, Language Testing 12 (1): 54-71.

[10] McNamara, T. F. (1996) Measuring second language performance. Harlow: Longman. Rob Schoonen, Margaretha Vergeer and Mindert Eiting. (1997) The assessment of writing ability: Expert readers versus lay readers, 14 (2): 157—184.

[11] Shi L. (2001) Native-and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. Language Testing, 18, 303-325.

[12] Ying Zhang and Catherine Elder. (2010) Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary construct?, Language Testing, 28 (1):31—50.

[13] Youn-Hee Kim. (2009) An investigation into native and non-native teachers' judgements of oral English performance: A mixed methods approach, Language Testing, 26 (2): 187—217.

[14] Wang Haizhen. (2007) Jiyu Pingfen Guocheng Zhengju de Yingyu Zhuanye Siji Koushi Xiaodu Yanjiu (Validation of TEM4-Oral: Evidence from Raters' Assessment Process). Jiefangjun Waiguoyuxueyuan Xuebao (Journal of PLA University of Foreign Languages). Vol. 30 No. 4.

[15] Zhang jie. (2009) Exploring rating process and rater belief –Seeking the internal account for rater variability, Unpublished Phd Dissertation, Guangdong University of Foreign Studies, China.

[16] Zhang Jie & He Lianzhen. (2008) Study of Sources of Score Variability in Performance Assessment Using MFRM: A Case of Speaking Test in Pets Band 3, CELEA Journal (Bimontyly), Vol. 31 No. 4.

[17] Zeng Yongqiang. (2009) The computerized oral English test of the National Matriculation English Test. In Cheng Liying & A. Curtis (Eds.), English Language Assessment and the Chinese Learner (pp. 234-247). New York and London: Taylor & Francis.