English Language Teaching
www.ccsenet.org/elt
Vol. 3, No. 1
March 2010

On Rater Agreement and Rater Training

Binhong Wang
School of Foreign Languages, Harbin Institute of Technology
PO box 443, 92 West Dazhi Street, Nangang District, Harbin 150001, China
Tel: 86-451-8641-3827 E-mail: elizawbh@yahoo.com.cn

Abstract

This paper first analyzed two studies on rater factors and rating criteria to raise the problem of rater agreement. After that the author reveals the causes of discrepencies in rating administration by discussing rater variability and rater bias. The author argues that rater bias can not be eliminated completely, we can only reduce the error to a cetain degree by training raters. The study on rater factors can help us better understand rater variability and rater bias, train raters more effectively and find out ways to modify the scores given by raters. The author suggests that rater files which contain rater information including each rater bias tendency should be established and kept so that information can be retrieved about the selection of raters, the interpretation as well as the modification of the scores given by raters. Raters need to receive pre-service training, on-service training and pilot -on-task training.

Keywords: Inter-rater reliability, Rater factors, Rating scale, Rater training

1. Introduction

Because of the individualized uniqueness and complexity, either the assessment of written discourses or oral speeches has long been a tough issue to raters. Therefore reliability and validity have been the focus of study in the field of performance assessment. Researchers have recognized that rater judgments have an element of subjectivity, and rater judgments of the same writers or speakers often vary. As to inter-rater reliability, historically, studies used to focus on inter-rater reliability which is calculated statistically. And studies have shown that a high inter-rater reliability is not hard to get (Shohamy, 1983; Fulcher, 2003). However, as to the interpretation and application of the assessing criteria, some scholars point out that raters differ in both the interpretation and application of the criteria (Bachman, 1990). According to some researchers, even training can not eliminate the above differences (Lumley & McNamara, 1995). Therefore, since the 1990s, studies of inter-rater reliability have shown a tendency to shift from a quantitative study to a qualitative study which focuses more on the process of assessment. During this period, “think-aloud” and “immediate retrospection” have been adopted as research methods (Cumming 1990; Cumming et al, 2002; Lumley, 2002; Orr, 2002). Raters are asked to provide retrospective written reports as well as introspective verbal reports on their thought processes. Thus, the rater, as opposed to the test materials, candidates or rating scale is used as the window through which the evaluation of second language speaking performance can be observed. This type of qualitative data can tell us about the process of discriminating between candidates (Pollitt and Murray, 1996), and is useful for the purpose of rater training (Weigle, 1994), and for investigating how raters reach their decisions (Milanovic et al., 1996; Taylor, 2000). The latest research of the type in China is a study on the interpretation and application of the assessing criteria in TEM4-Oral (Wang Haizhen, 2008). The research adopts a qualitative research method based on simultaneous reports and recall reports (immediate retrospection) as well as simulated recall.

This paper aims to raise the problem of rater agreement. By analyzing studies on rater agreement and discussing rater variability and bias, the author intends to come up with ways to compensate for the systematic error caused by rater factors.

2. Studies on Rater Agreement

A representative research on rater agreement is a qualitative study on raters’ interpretation and application of the Rating Criteria in TEM4-Oral (Wang Haizhen, 2008). In this experiment, 24 raters with 4 males and 20 females, who came from 11 different universities in China, were studied. All the raters had English Major teaching experience which varied from 1 year to 21 years. Rater’s age ranged from 26 to 42. When it comes to rating experience, 11 of them were newly trained raters TEM4-Oral, while 13 of them had TEM4-Oral rating experience ranging from 1 year to 9 years. In other words, raters differed in their backgrounds. To study the decision-making process of rating, qualitative research methods of think-aloud and immediate retrospection were used. The experiment began on December of 2005. The rating experiment was composed of three stages: preparation, rater training and rating. Preparation aimed to make raters get familiar with TEM4-Oral as well as the rating criteria. Rater training consisted of the essential training in order to be a qualified TEM4-Oral rater as well as subsequent retrainings to maintain the qualification. Rating involved asking each rater to rate independently on 5 chosen sample examinees (i.e. 5 tapes which covered 4 levels were chosen among one

group of 32 tapes obtained in 2005 TEM4-Oral). The raters were asked to rate on task 3 only, a 3-minute instaneous speech, judging by the content, sounds and intonation, vocabulary and grammar. The study finally came to three conclusions:

1). Raters applied not only the given rating criteria but also other criteria as well.

2). Different raters interpreted and applied the rating criteria in different ways.

3). There were discrepencies in the ratings of different raters toward the same examinee.

But the study also showed that despite the discrepencies in the rating process, this experiment had a high interrater reliability of 0.972.

Since we have come to the conclusion that raters show differences in interpreting and applying rating criteria, and now if we go a step further, we may presume that native raters and nonnative raters may show even greater differences in rating. And this is proved by a study on the differences in native and nonnative judgements of Chinese contestants' performances in an English speaking contest (Wen Qiufang, Liu Xiangdong & Jin Limin, 2005). The data for analysis was from the Semi-final of 2004 "CCTV Cup" English Speaking Contest in which there were 96 contestants and 11 judges, including five native speakers and six nonnative speakers. The native judges were foreign teachers teaching English in Chinese tertiary institutions while the nonnative judges were English professors whose mother tongue was Chinese. The results showed that native and nonnative judges did show significant differences in their average scores for the 96 contestants. However, such differences did not affect the outcome of the contest concerning 75 % of the semi-final winners. For the remaining 25 % of the semi-final winners upon whom native and nonnative judges disagreed, native judgments weighed more than nonnative ones. The interviewing data indicated that native and nonnative judges differed principally in how they rated the linguistic forms and the content of the contestants' oral performance. Native raters and nonnative raters had apparent different rating criteria with different focuses. Native raters attached more importance to the intelligibility and ignore those phonetic and grammatical mistakes which wouldn't affect comprehension while nonnative raters had low tolerance of linguistic mistakes. The study also found that native raters weighed coherence over relevance of the contestants' speeches while nonnative raters had low tolerance of speakers' irrelevance in the instaneous speech performance.

In this study, we can see that although native raters and non-native raters displayed apparent discrepencies in applying the rating criteria, the interrater reliability was still comparatively high with agreement on 75% of the semi-final winners. But we can also see native raters and nonnative speakers applied different rating criteria with different focuses. Next the paper will give interpretation to the discrepencies by discussing rater variability and rater bais.

3. Interpretation and discussion

From the two cited studies, we can see that, in a rating administration, raters may apply not necessarily the same criteria to the examinee and even when they apply the same criteria, their evaluations or judgments on the examinee's oral proficiency may differ. Thus inter-rater inconsistency is an unavoidable source of error. Douglas (1994:134) took it as a source of bias in language use and was less than optimistic about the chances of oral test raters being standardised, claiming that "It is almost axiomatic that no two listeners hear the same message".

The author argues that the systematic error caused by raters can never be eliminated completely, we can only reduce the error to a cetain degree by training raters. The study on rater factors can help us better understand rater variability and rater bias, train raters more effectively and find out ways to modify the scores given by raters.

3.1 Rater factors

3.1.1 Rater variability

Raters are not born raters even though they're native speakers. Rater factors, such as their mother tongue, age, gender, educational background, research areas, knowledge about ESL learning and oral ability development, personal character, experience as a rater, whether they have received any training to be raters, etc., will affect their ratings. Among the studies about rater factors, one pilot study was made on the effects of background characteristics of interviewers on the inter-rater reliability of the oral testing procedure for the Senior High School French Program in the Province of Newfoundland of Canada (Flynn, 1991). The research came to the following conclusions: (1) there were significant differences in the ratings of the oral interview; (2) these differences were related to the oral proficiency of the interviewer; (3) these differences were in the areas of assessment of vocabulary, grammar and to some extent fluency. The research also found that in rating the five oral proficiency factors (vocabulary, grammar, comprehension, fluency, pronunciation) respondents rated the most globally viewed factors, comprehension and fluency, as the most important factors. The findings were consistent with the findings of Higgs and Clifford (1982): grammar appears to be a more important factor for those raters whose language proficiency level is higher while vocabulary appears to be a more important factor for less proficient interviewers.

Generally speaking, rater variability can manifest in various ways (Bachman & Palmer, 1996; McNamara, 1996;

Lumley, 2005). Raters may differ (a) in the degree they comply with the scoring rubric, (b) in the way they interpret criteria employed in operational scoring sessions, (c) in the degree of severity and leniency exhibited when scoring examinee performance, (d) in the understanding and use of rating scale categories, or (e) in the degree to which their ratings are consistent across examinees, scoring criteria, and performance tasks. This paper will continue to discuss rater variability in terms of (c), in the degree of severity and leniency exhibited when scoring examinee performance.

3.1.2 Rater bias

The term rater bias refers to rater severity or leniency in scoring, and has been defined as 'the tendency on the part of raters to consistently provide ratings that are lower or higher than is warranted by student performances' (Engelhard, 1994:98). Numerous studies have been made on rater bias pattern which aimed to offer implications in rater training. Wigglesworth (1993:314) found that there was a reduction of rater bias and improvement in internal consistency in subsequent ratings after raters adjusted their ratings according to the feedback of rater bias analysis. However other researchers have found that rater variability persists in spite of extensive rater training and screening (Engelhard, 1992, 1994; Lumley, 2002, 2005; Lumley & McNamara, 1995; McNamara, 1996). Lumley and McNamara (1995:57) noted that although rater training reduces random error and makes raters more self-consistent, it cannot eliminate rater variability. McNamara (1996:127) questioned whether it is even necessary to have complete rater consistency.

The author argues here that rater bias can never be eliminated, because raters differ in age, gender, nationality, life experiences, educational background and research interests as well as personal character. They also differ in their experiences as raters. Therefore their ways of perception also differ. Towards the same utterance input, raters perceive with different focuses and from different perspectives, which lead to discrepancies in their assessing. In other words, raters' interpretation and application of the same rating scale may vary from rater to rater. So we can't require all raters to be absolutely uniform in interpreting and applying criteria and we can't expect to get absolutely the same assessing results from all the raters. An elaborate rating scale may help reduce such differences but can not eliminate it completely. As long as the final scores given by each rater leading to similar assessing results which are within discrepancy tolerance, we can tolerate the discrepancies and still expect to achieve a high interrater reliability. Trained raters can arrive at similar or comparable conclusions/assessments through different emphases. One may concentrate on grammar, another on fluency – or different raters may have dissimilar conceptualizations of fluency. This is not problematic, especially given that the subcomponents of holistic scales are highly correlated. As mentioned before, in case one, the interrater reliability is 0.972 and in case two, the percentage of total agreement on semi-final winners is 75%.

3.2 Rater files and rater training

The significance of the study on rater agreement is to call for studies on rater factors, rater bias and rater training. The study of rater factors and rater bias can give us some clues of the discrepancies in assessing and provide enlightening ways to deal with or compensate for rater bias.

In this paper, the author proposes that rater files which contain rater information including rater bias tendency should be established and kept so that information can be retrieved about the selection of raters, the interpretation as well as the modification of the scores given by raters. For example, in this case study, we can establish a general rater file as follows: see Table 1 (Note: here the personal information of Rater 1 is presumed)

In Table 1, Rater bias tendency coefficient indicates whether a particular rater's rating is above or below the average rating as well as the degree. Suppose in this case study, R1's rater bias tendency coefficient is "-1", which means 1 point below the average score. R3's rater bias tendency coefficient is "+1", which means 1 point above the average (suppose the total score is 5-point). Accordingly, Rater 1's modification coefficient is "+1", which means that the modified score is obtained by adding 1 point to the original score. Rater bias tendency can be found by having raters make assessments on several cases which are representative in terms of examinees and peformance tasks. Suppose this is a general rater file. Considering that rater bias tendency may vary with particular assessing tasks, examinees, etc. therefore, the author advises pilot-on-task assessings before raters assess a particular performance test. Based on the pilot on-task assessings and the general rater file, rater bias tendency and score modification strategy can be worked out.

Another advantage of pilot on-task assessing is that raters can get feedback from the rating analysis and adjust their rating (Wigglesworth, 1993:314) or cloning raters (Alderson, 1991:64). If a rater's pilot assessing is beyond the discrepancy tolerance, we can screen out that rater in the assessing mission.

Rater training cannot eliminate rater bias, but can only make raters more self-consistent. Studies showed that results of rater training may not endure for long after a training session, so the practice of holding a moderation session before each test administration is necessary to allow raters to re-establish an internalized set of criteria for their ratings. Analyses confirm that judge differences survive after trainining, so it seems that at every administration, new calibrations of rater characteristics are required (Lumley & McNamara,1995). Rater training is not a once-for-all matter, it is on-going business. (see appendix)

4. Conclusion and suggestion

Based on the above analyses and discussions, we come to the following conclusions:

1) A high inter-rater reliability is not difficult to be achieved.

2) There are apparent discrepencies in raters' interpreting and applying the rating criteria. Rater bias exists and rater factors need to be studied.Rater bias can never be eliminated. Rater bias analysis and rater training can only help reduce the degree of discrepancies.

3) Rater files are helpful in seleting raters, interpreting rating scores and modifying scores.

4) Raters need to receive pre-service,on-service training constantly to maintain their qualification as raters.

Thus this paper gives the following suggestions:

1) The research on oral English testing and assessing need to receive increasing concern in the future.

In 2007, the Education Ministry of China issued new College English Teaching Requirements, which states that the teaching objective of College English teaching is to develop students' integrated English abilities, especially listening and speaking ability. But researches on oral English teaching and assessing are still limited in China, especially to non-English majors. Future research should include oral English teaching, textbook compiling, and increasing the reliability and validity of large-scale oral English testing (Wang Lifei and Zhou Dandan, 2004:8).Therefore, it can be predicted that the study on oral English teaching, testing, assessing and rater training will be in demand in universities of science in China.

2) Establish rater files .

Rater factors and strategies to deal with rater bias need to be studied. Establish rater files which collect rater information concerning age, gender, educational background, research areas, experiences as raters, and especially their rater bias tendency (above or below the average assessing, if possible, the degree of this tendency). Studies should be carried toward strategies to reduce or compensate for rater bias.

3) Raters need to be trained constantly.

Potential raters need to be trained before they are assigned to be raters. To ask potential raters to make assessment on some typical cases is an effective way. Raters need to be trained in terms of applying holistic rating method and an analytical rating method. In the training process, simultaneous reports (think-aloud), immediate retrospection as well as stimulated recall can be adopted. Raters need to be trained constantly and new raters should be trained with experienced raters together to gain some valuable experience. Besides, it's highly advised that pilot on-task training should be carried out before each assessing task. The pilot on-task training can help conform raters to the rating criteria and reduce rater bias, to some extent, cloning raters. In pilot assessing, those potential raters whose ratings are beyond discrepancy tolerance should be considered unqualified for the specific assessing task and be screened out.

4) Raters' making justifications should be incorporated into rater training course.

Raters' making justifications is a helpful and essential way to learn about the process of raters' interpreting and applying the rating scale.

Appendix

Rater training at a large North-western American university consists of the following:

1) Completion of a 20 hour rater training program -- raters must achieve 70% agreement overall to move to the next phase, repetition with new sets until 70% is achieved -- 90% of trainees meet the 70% requirement after 20 hours

2) Rating in the August admin as a apprentice rater -- 80% must be achieved -- if achieved the rater is certified, if not rater continues as a practice rater until 80% agreement achieved

3) Regular rater training meetings every month

4) Summer calibration 20 hours every year

References

Alderson, J.C. (1991). Dis-sporting life. Response to Alastair Pollitt's paper: 'Giving students a sporting chance'. In J.C. Alderson & B. North (Eds). *Language Testing in the 1990s*: The Communicative Legacy. Hemel Hempstead: Modern English Publications in association with the British Council.

Bachman, L.F. (1990). *Fundamental Considerations in language Testing*. Oxford: Oxford University Press: 177-181.

Bachman, L.F. and Palmer, A.S. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Cumming, A. (1990). Expertise in evaluating second language composition. *Language Testing* 7:31-51.

Cumming, A., R. Kantor, & D.E. Powers. (2002). Decision making while rating ESL/EFL writing tasks: A Descriptive framework. *The Modern Language Journal* 86/1:67-96.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing* 11: 125–144.

Engelhard, G. Jr.(1992). The measurement of writing ability with a many faceted Rash model. Applied Measurement in Education 5/3:171–191.

Engelhard, G. Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement* 31/2: 93–112.

Fulcher, G. (2003). *Testing Second Language Speaking*. London: Pearson Education Limited.

Flynn, Kevin Francis. (1991). A Pilot Study of the Effects of Background Characteristics of Interviewers on the Inter-rater Reliability of the Oral Testing Procedure for the Senior High School French Program in the Province of Newfoundland [D].Memorial University of Newfoundland (Canada).

Higgs, Theodore, V. and Ray Clifford. (1982). "The Push toward Communication." In Charles James (Eds), *Curriculum, Competence, and the Foreign Language Teacher.* Lincolnwood, Illinois: National Textbook Company, pp.55-78.

Linacre, J. M. (1989). Rasch models from objectivity: A generalization. Paper presented at the International Objective Measurement Workshop, Berkeley, CA.

Lumley, T. & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12/1: 54–71.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to raters? *Language Testing* 19/3: 246-276.

Lumley, T. (2005). *Assessing second language writing*: The rater's perspective. Frankfurt am Main: Peter Lang.

McNamara, T. F. (1996). *Measuring second language performance*. New York: Longman.

Milanovic, M., Saville, N. & Shen, S. (1996). A study of the decision-making behavior of composition markers. In: Milanovic, M., Saville, N. (Eds.), Studies in *Language Testing 3*: Performance Testing, Cognition and Assessment. Cambridge University Press, Cambridge.

Orr, M. (2002). *The FCE speaking test*: Using rater reports to help interpret test scores. System, 30/2: 143-154.

Pollitt, A. & Murray, N.L. (1996). What raters really pay attention to? In: Milanovic, M., Saville, N. (Eds.), *Studies in Language Testing 3*: Performance Testing, Cognition and Assessment. Cambridge University Press, Cambridge.

Shohamy, E. (1983). "Interrater and intrarater reliability of the oral interview and concurrent validity with cloze procedure in Hebrew". In *J.W.Oller* (ed.). Issues in Language Testing Research. Rowley, MA: Newbury House.

Taylor, L. (2000). *Approaches to rating scale revision*. EFL Research Notes 3:14–16.

Weigle, S. (1994). Effects of training on raters of ESL compositions. *Language Testing* 11: 197–223.

Wen Qiufang, Liu Xiangdong & Jin Limin.(2005). Native and Nonnative Judgements of Chinese Learners' English Public Speaking Ability. *Foreign Language Teaching and Research* 37/5:337-342

Wang Haizhen. (2008). A Study on Raters' Interpretation and Application of the Rating Criteria in TEM4-Oral. *Theory and Practice of Foreign Languages Teaching* 2:33-39.

Wang Lifei & Zhou Dandan. (2004). On 12 Years' Research of Oral English in China: Retrospect and Current Situation. *Foreign Language World* 6: 7-14.

Wigglesworth, G. (1993).Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction.

Language Testing 10: 305-335.

College English Teaching Requirements (Revised). (2007). http://www.chinanews.com.cn/edu/kong/news/2007/09-26/1036802.shtml

Table 1. General rater file

Rater ID	Rater Name	Gender	Age	Mother tongue	Educational background/Research area	Years' of being raters	Rater bias tendency coefficient	Score Modification Coefficient
R1		Female	30	English	PhD in Applied Linguistics/English teaching methodology	7	-1	+1
R2		……						
R3…….								