Rater Reliability on Criterion-referenced Speaking Tests in IELTS and Joint Venture Universities

By James Lee

Lee, J. (2014). Rater reliability on criterion-referenced speaking tests in IELTS and Joint Venture Universities. *English Teaching in China*, 4, 16-20.

Introduction

The global expansion of Higher Education has been particularly marked in China, where a number of English-medium universities have been established over the past decade. These institutions, generally known as Joint Venture (JV) Universities, tend to offer a significant amount of language support through intensive English for Academic Purposes (EAP) provision, and there is a need, as on most courses, to accurately assess the language level of students. Speaking assessment constitutes a significant amount of such testing and is of great importance for both the students and the institution, as it provides evidence of both the students' abilities and the effectiveness of the teaching.

While JV assessment is in-house, the International English Language Testing System (IELTS) is a high-stakes global test which uses assessment criteria as a basis for its direct assessment of candidates' spoken English. Indeed, Alexander, Argent and Spencer (2008, p. 327) have claimed that large-scale language exams, such as IELTS, introduced the concept of assessment criteria and are responsible therefore for improvements in reliability within EAP institutions. The IELTS test is particularly influential in China because of the sheer number of candidates who sit it every week.

Historically, there was reluctance to use subjective tests due to their perceived unreliability (Hughes, 1989, p. 53). Currently, the use of criterion-referenced speaking assessments is widespread. One issue which they raise, however, is that of rater reliability. Here rater reliability will be considered; in particular, to what extent raters' interpretations of criteria can be considered to be reliable. The article will then examine particular reliability issues faced by IELTS and JV Universities and identify ways in which those challenges can be met.

Reliability

An effective test needs to be reliable. Reliability ensures that the score is an accurate reflection of the student's knowledge. Scoring should be consistent no matter when or where the test is taken (test reliability), nor who marks it (rater reliability). Reliability, therefore, is an assurance that test results are the best possible indicator of a candidate's performance. It implies consistency and accuracy of measurement.

There are generally considered to be two aspects to rater reliability. Inter-rater reliability is concerned with the consistency of scoring displayed across a group of raters, whereas intra-rater reliability refers to how consistent

one rater is at giving an accurate score on different occasions (Bachman, 1990, p. 178). Inter-rater reliability is established by comparing the scores of different raters, while checking intra-rater reliability involves re-marking tests after a long period of time. There appears to be no easy way to ensure rater reliability however. Hughes (1989, p.36) has claimed that rater reliability can never be guaranteed when there is a subjective element to the rating.

Factors affecting rater reliability

Different areas of inter-rater inconsistency have been identified. Certain raters can be classed as being more severe or more lenient than the norm. This may stem from the rater's personality or cultural background. A rater's particular rating style means that they do not divide their attention evenly among the various criteria used. Carey, Mannell, and Dunn (2011) found that familiarity with a candidate's accent was likely to lead to an examiner awarding higher marks for pronunciation in speaking tests, and non-native examiners rated candidates from their own country higher than those from different locations. Similarly Winke, Gass, and Myford (2013) have claimed that if an examiner has experience of communicating in the L1 of the test-taker, they are more likely to show leniency to that candidate. This would seem to support the view of McNamara (2000, p. 38) that the score cannot be separated from the rater; it is a reflection both of the performance of the candidate and the beliefs of the scorer. Subjectivity causes even experienced raters to disagree over borderline cases. As a result, whether a candidate achieves a certain score can be dependent on how lucky (or unlucky) they are in being assigned a certain examiner.

Inconsistency in inter-rater reliability also stems from the criteria used for rating; raters understand and use the rating scales differently, or have conflicting interpretations of the criteria themselves. Raters need to arrive at a set of scores by considering their overall impression of a candidate's answers, noting any specific features contained within, and matching these to the wordings of the descriptors. However, no set of descriptors can adequately cover all of the possible language produced – raters may need to develop a range of strategies in order to help negotiate these problem areas, which can lead to a conflict between their intuition and the criteria. It is unclear how raters resolve this tension. Criteria need to be as wide-ranging as possible, but although broad criteria allow raters a greater degree of flexibility when it comes to assigning marks to a candidate, they also necessitate a significant degree of judgement on the part of the rater. According to Alderson, Clapham, and Wall (1995, p. 108), it is a considerable challenge for raters to understand the principles behind rating scales and to interpret them consistently. For Alexander et al. (2008, p. 335) it is the complexity of texts which mean that it is often very difficult to interpret assessment criteria. Speech can be unbalanced, strong in places, weak in others, which creates confusion in the mind of the rater. It is crucial for raters not to be constrained by the descriptors but to use them to justify their decisions (ibid.).

Intra-rater reliability

Interpretation of scoring criteria can also affect intra-rater reliability. There is a possibility that when rating a number of examinees who make grammatical errors, after a while a rater may begin to focus more critically on the grammar descriptors, and candidates who were marked first will score higher than later ones. Alternatively scoring may be relaxed; for example, the same pronunciation issues heard again and again may seem less explicit in later candidates compared to previous speakers (Bachman, 1990, p. 179). This may be linked to marking load. In a test where students are being judged on content, a rater who has heard the same ideas up to 80 times may feel that later test-takers lack originality.

Rater reliability in the IELTS test

Over 1.7 million IELTS tests are taken every year, 300,000 of which are sat in China (British Council, n.d.). The Speaking component is a criterion-referenced test, rated by a trained Examiner. Detailed performance descriptors measure different aspects of candidates' spoken capabilities, and the exam uses analytic rating scales, which allow the rater to narrow down the focus to one or two grades, and then

'fine-tune' their score. IELTS has funded a number of research projects on rater reliability, and made them publicly available on its website in an effort to ensure that the scoring procedure is fair and rater reliability is high. Blackhurst (2004, p. 18) states that when speaking scores were double marked in 2003, there was a correlation of 0.91, which is considered a high level of rating reliability. Nevertheless, Uysal (2010, p. 316) has suggested that the claim that use of analytic rating scales means that reliability increases is unsubstantiated.

IELTS requires their Examiners to examine on a regular basis, and Examiners need to re-certify every two years, or if they go three months without examining. In China, examiners may rate up to 20 speaking exams per day, for as many as four or five days in a row. This heavy testing load may have implications on the intra-rater reliability of the test, particularly if this is a 'second job' for the rater. Uysal (2010, p. 319) has emphasized the need for constant calculation of both intra- and inter-rater reliability measures. The vast numbers of tests being taken every week will similarly have an effect on inter-rater reliability. McNamara (2000, p. 42) has also identified potential problems with the IELTS reference to "native-speaker level competency" as a standard to be achieved. He suggests that this is a misleading term, as performances of native speakers (NS) vary considerably. A number of non-native speakers (NNS) are employed as IELTS Examiners, but only after achieving a Band 9 ('native speaker level') in the test themselves. To date, however, there is scant research on differences in rating between NS and NNS Examiners in IELTS.

Hall (2010, p. 324) claims that the IELTS marking and standardization process is as rigorous as possible. The IELTS website (IELTS, n.d.) states that "IELTS Examiners undergo intensive face to face training and standardization to ensure that they can apply the descriptors in a valid and reliable manner". Raters take a two-day training course prior to qualification as a Speaking Examiner and at the end of this course, are required to rate a number of candidates accurately in order to progress to Examiner status. The training procedure uses what Alderson et al. (1995, p. 131) refer to as "reliability scripts". These are scripts (or, for the speaking component, videos of exam performances) for which a consensus on the score has been reached. Trainees should identify features in the criteria which explain the scoring. Newly-qualified Examiners also have their first assessments double-marked by an experienced Examiner Trainer. Raters are then randomly moderated by Examiner Trainers, however perhaps as little as 1% of their examining will be moderated in this way in China. Thus the vast majority of scores are single-marked, i.e. trusted solely to the judgment of that examiner. The importance of double-marking is a complex and contested issue. Uysal (2010, p.315) claims that it is widely accepted that scoring accuracy can be improved by multiple marking, yet Hall (2010, p.323) highlights problems which can arise when tests are double-marked, for example, if two raters arrive at the same score, we cannot be sure that both arrived there for the same reason.

Rater reliability in Joint Venture Universities

JV Universities in China provide EAP instruction to students on an unprecedented scale. At Xi'an Jiaotong-Liverpool University (XJTLU) in Suzhou, for example, there are over 2000 new students each academic year, and each has up to 10 hours per week of English classes and needs to achieve a pass grade to progress in their undergraduate studies. JV institutions need to accurately measure the Academic English ability of significant numbers of students, across a number of academic disciplines. As a result, ensuring rater reliability is crucial. The large numbers of test takers involved present various problems:

1) Tutors tend to have a large amount of rating to do in a very short time, which can affect intra-rater reliability;

2) Heavy workloads reduce opportunities for double marking;

3) Differing interpretations of criteria are multiplied as the number of raters increases, affecting inter-rater reliability;

4) JV raters tend to come from a variety of backgrounds. A small study by Shi (2001) found that Chinese nationals and Native

Speakers justified their ratings in different ways and were unsure how to apply different criteria in writing tasks. This finding could reasonably be extended to speaking tasks. A lack of consensus among teachers is likely to prove confusing for students both in terms of teaching and assessment feedback;

5) One final important issue is whether tutors should assess their own students or another teacher's classes.

A key way to improve rater reliability is to hold regular standardization meetings. A number of researchers have emphasized the importance of these meetings. Alderson et al. (1995, p. 130) propose that ideally there should be a Chief Examiner who conducts standardization meetings. It would appear to be extremely important that in JV Universities there is at least an Exam Officer role, with overall responsibility for training raters on how to interpret criteria, the importance of which has been emphasized by Hughes (1989, p. 55). Crucially, standardization allows discussion of discrepancies in scores given by different raters, with a particular focus on the way that descriptors are interpreted by individual raters. McNamara (2000, p. 44) suggests that peer pressure is useful for 'reining in' some more extreme markers. Monitoring of raters is obviously necessary to ensure that institutional assessment standards are being complied with. While the judgment of a candidate's performance is by nature multi-faceted and complex, and we can never be entirely sure how an original marker arrived at their score, sufficient training and reorientation, with a particular emphasis on dealing with conflicting interpretations, can help improve reliability and familiarity with how institutions view the criteria of their tests. Standardization sessions provide an opportunity for teachers to explore their cultural backgrounds and compare rater expectations, and may give insights into educational systems and teacher beliefs. As new institutions, it is important for JV Universities to be clear and transparent in their assessment procedures.

Alexander et al. (2008, p. 335) state that the assessment criteria need to be used as a framework for discussion in standardization meetings. This ensures consistency as it allows markers to deal with discrepancies collaboratively, and evaluate the effectiveness of the descriptors. This collaboration enables new teachers to become familiar with the assessment practices and expectations of an institution and can lead to improvements of the existing criteria. Such meetings may lead to frank exchanges and a degree of, hopefully minor, conflict, yet this is a necessary by-product of employing independent thinkers who consider texts deeply and react differently to them. Evidence from the recruitment practices of JV Universities in China thus far seems to suggest that the importance of employing well-qualified teaching practitioners is being recognized and this should continue in order to further intelligent debate centred around speaking test descriptors.

Final thoughts

Rater reliability is an important touchstone of how effective a speaking test is at measuring what it sets out to measure. If rating is not reliable, all of the hard work invested in constructing a valid test will have been in vain (Alderson et al., 1995, p. 105). The use of criterion-referenced assessment offers the chance to measure a candidate's test performance in detail. Despite this, the interpretation of those criteria by the person assigned to rate a candidate will always have an element of subjectivity. While this may suggest that reaching consensus among raters is an impossible goal, this paper has identified certain steps toward improving reliability.

IELTS tries to ensure the reliability of its tests by putting a large emphasis on initial examiner training, using reliability scripts to bring new examiners into line with the expectations of the organization; examiners are closely monitored to begin with, and thereafter obliged to examine on a regular basis.

However, regular practice may not be enough to ensure reliability, particularly in cases where examiners have a heavy workload. The rapid expansion of JV Universities means that they will need to address similar concerns about rater reliability in order to prove the validity of their assessment. It seems that it would be particularly useful to hold standardization meetings to explore what descriptors represent and to understand

clearly how raters interpret and apply these criteria. Close monitoring and analysis by experienced raters is also essential. JV Universities have a large amount of expertise to draw upon, and this should be channeled effectively to ensure that standards are continually being met.

References

Alderson, J., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: CUP.

Alexander, O., Argent, S., & Spencer, J. (2008). *EAP essentials: A teacher's guide to principles and practice.* Reading: Garnet.

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford: OUP.

Blackhurst, A. (2004). IELTS test performance data 2003. *Research Notes, 18,* 18–20.

British Council (n.d.). *IELTS.* Retrieved from http://www.britishcouncil.org/china-exams-ielts.htm

Carey, M., Mannell, R., & Dunn, P. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, *28*, 201–219.

Hall, G. (2010). International English Language testing: A critical response *ELT Journal, 64*(3), 321–328.

Hughes, A. (1989). *Testing for language teachers.* Cambridge: CUP.

IELTS Homepage (n.d.). Retrieved from: http://www.ielts.org/researchers/score_processing_and_reporting.aspx

McNamara, T. (2000). *Language testing.* Oxford: OUP.

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

Uysal, H. (2010). A critical review of the IELTS writing test. *ELT Journal, 64*(3), 314–320.

Winke, P., Gass, S., & Myford, C. (2013). Rater's L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*, 231–252.

James Lee is an EAP Tutor at XJTLU in Suzhou , China. He has taught in Slovakia, Japan, Taiwan, Korea, Mexico and the UK. He is currently conducting research into International Students' experiences studying EAP in China.

james.lee@xjtlu.edu.cn.