Universe Scientific Publishing

DOI: 10.18686/ahe.v6i9.4729

Comparison of TOEFL iBT and ESOL in Reliability

Zixin Li

School of Foreign Studies, Xinjiang University, Xinjiang Urumuqi City, 830000

Abstract: TOEFL iBT and ESOL are two language proficiency tests, and they are widely accepted by test takers worldwide. Both tests assess test takers' speaking skills, but they vary to some extent. This study will compare these two tests in terms of reliability criteria.

Keywords: TOEFL iBT; ESOL; Rater reliability; Test reliability; Comparison

1. Introduction

TOEFL iBT (Test of English as a Foreign Language Internet-based test) and ESOL (English for Speakers of Other Languages) are two language proficiency tests that include speaking, writing, reading, and writing skills, they are widely accepted by test takers worldwide. TOEFL iBT test is to measure a test taker's English ability at the university level (Ets TOEFL, n.d.), and the TOEFL speaking component will be a measure of oral communication in an academic context. Trinity's ESOL Skills for Life qualifications are designed to meet the demands of learners who have chosen to make a life in the UK (Trinitycollege, n.d.). ESOL speaking test contains elements of task-based assessment for the needs to settle in the country and get by on a day-to-day basis (Trinitycollege, n.d.). TOEFL iBT and ESOL both have speaking components, however, their speaking tests may not only have similar aspects but also distinguish each other in various aspects.

Bachman and Palmer's framework (1996) of the test is influential and worth considering when designing and evaluating a test (Newton & Shaw, 2014), they propose test usefulness includes reliability, construct validity, authenticity, interactiveness, impact, and practicality (Bachman & Palmer, 1996). Validity and reliability as two major concerns and necessary qualities of measurements in the speaking test (Bachman, 1990; Bachman & Palmer, 1996; Alderson et al, 1995), have been discussed a lot. Though qualities cannot be analysed and evaluated independently and should be considered together, because of limited space, I would compare TOEFL and ESOL speaking tests mainly based on the criteria of reliability.

2. Comparison

Reliability is the basic criteria for a test to adhere to ensure the score accurately reflects a student's knowledge (Lee, 2014). Reliability refers to the score consistency on a test no matter what situation happen (Bachman & Palmer, 1996). But test performance would be impacted by unrelated factors other than key language ability we aim to measure (Bachman, 1990, p.159). Hughes (2003) further proposes that the scores would likely to be different if one test had been taken at a different time or on a different setting as human do not always behave same on every occasion, thus we cannot completely or easily trust any test scores. According to their views, if scores share higher similarity and fewer unrelated factors involved, then the test scores and even the test itself would be more reliable (Hughes, 2003; Bachman, 1990). Scoring should be consistent regardless of different test conditions (test reliability), or different markers (rater reliability) (Lee, 2014). Concerning with the rater reliability and test reliability, I would argue that TOEFL iBT speaking test is more reliable than ESOL speaking test in rater reliability, but both of them have unreliable factors in test reliability.

3. Rater reliability

An oral test relies much on interaction and communication, it may be greatly influenced by rater reliability, while rater reliability can be classified into inter-rater reliability (whether different raters can give consistent judgement on one test taker of the same phenomenon) and intra-rater reliability (whether an individual rater can produce consistent measurements after a period) (Gamaroff, 2000).

First, training scorers is an important way to ensure rater reliability. The oral interview is subjective, for the reason that lacking correct criteria for correctness, judgement is just founded on a rating scale (Shohamy, 1983). Besides, scorers' have variety, subjectivity and different experience, even raters are provided with guidance and samples, they may still have different interpretations of the rating scale in the process of scoring (Shohamy, 1983). However, training as Luoma (2014, p.192) describes, usually contains an introduction to the test and its rating criteria, explanation of rating scale and its different levels, and the former rating examples may also be shown. Scorers may have the opportunity to practise rating after being shown with the examples, they may also have chances to report their scores and discuss the reasons with other trainees. All these procedures would help trainees learn to apply these criteria according to the system's conventions. Training scorers enable trainees to be familiar with the rating scale and scoring procedure (Shohamy, 1983), and after the uniform training, raters may be taught to evaluate performance in the systems' terms. In principle, raters may reach an agreement in terms of the scoring criteria through training which may increase inter-rater reliability to some extent.

In ESOL, all Trinity examiners and markers must complete regular training and standardisation, they would be monitored regularly to ensure Trinity's standards are maintained (Trinitycollege, 2013). In TOEFL, raters are trained extensively, they have a strict standard and training mechanism; trainees have to use the criteria correctly in the training and pass oral scoring certification first, then they would have a chance to become a qualified scorer (TOEFL® Research Insight Series, n.d.). Besides, raters are calibrated daily, the calibration includes task familiarization, guidance on scoring the task, and practice on a range of responses (Ets TOEFL, n.d.). Before scorers take part in every live scoring session, they must pass a topic-specific calibration test beforehand; they are not allowed to score unless they pass this calibration test (TOEFL® Research Insight Series, n,d.). In addition, during each scoring session, raters are monitored and supervised by scoring leaders. If any problems arise, raters are retrained or replaced(TOEFL® Research Insight Series, n,d.). All these procedures can greatly ensure the training qualities and effectively avoid scorer' bias.

In this sense, examiners in TOEFL iBT and ESOL speaking tests have gone through a qualification procedure and reached the basic reliability as scorers, but from the information mentioned above, I would assume TOEFL has clearer scrutiny than ESOL has.

However, rater reliability is also concerned with reconciling subjectivity and objective precision (Gamaroff, 2000). Training cannot guarantee the examiners will mark as they are supposed to (Alderson et a., 1995). In Orr's research (2002), raters may not always adhere to the rating criteria, they may unconsciously heed many aspects of the performances not relevant to the assessment criteria (Orr, 2002, p.153), such as a test taker's body language and eye contact. Thus, it is unwise to rely a test score solely on one examiner's judgement.

In ESOL speaking test, it is an unreliable factor, test takers are rated by only one rater, basically, most scores are single-marked, then score trusted solely to the judgment of that examiner, if the examiner is unconsciously affected by other factors, the scores would be unreliable. In this case, ensuring scorers' marks are reliable require double-marking in every part of the exam (Alderson et a., 1995), the reliability can be improved by correlating the marks given by two or more raters on the same students. Involving two scorers to double mark an examinee has achieved in TOEFL exams, its speaking items are scored not only by a network of a human judge using a rating scale but also by AI system. AI scoring and multiple human raters who do not know test takers' identities, would score candidates performance together, and "this way can prevent examiner's bias that can occur in other tests that use a face-to-face interview with a single rater"(TOEFL® Research Insight Series, n,d.).

4. Test reliability

Considering the test formats in ESOL, it is tested in the presence of an examiner who gives scores using a pre-established rating scale. It is a face-to-face oral interview, and the exam comprises two components: a 14-minute one-to-one conversation with a Trinity examiner, and a 15-minute discussion with three candidates, facilitated by a Trinity examiner. The format in ESOL speaking test focuses on the real face-to-face interaction between the interviewer with the interviewee and between two candidates. Though Norton (2005) argues that paired speaking format provides a relaxed atmosphere just as the same as a class situation for students, I don't agree with her. Likely, the content and style of the interviewer or the interaction between the interviewee and interviewer would affect the rating. In a group discussion section, the candidates may also influence mutually. Weir (2005) says that one person performance in the co-constructing assessment may affect another individual's performance, if one person is talkative, the other one is introverted, the introverted one may feel intimidated; If the two individuals have different language abilities, it may also be difficult for the other person to perform to his/her potential. Brown (2003) further adds that "if the candidates have different proficiency levels, the one with higher proficiency tends to get frustrated whilst the one with lower proficiency level feels subdued or anxious". In addition, O'Sullivan (2002) says the candidates will get higher marks if they share the same mother tongue. Not to mention if one candidate has a strong accent, it will cause more stress or anxiety on another candidate, which would accordingly produce unexpected performance, these factors would all influence the test reliability. By the way, it is also possible that a speaker of a lower level get a grade he does not deserve, such as the rater shows sympathy for the nervous or weaker candidate (Orr, 2002), then the scores would lose their reliability.

In TOEFL, it is a semi-direct format. "Semi-direct" is explained by Clark (1979, p.36) to describe those tests that are characterized "using tape recordings, printed test booklets, or other 'non-human' elicitation procedures, rather than through face-to-face conversation with a live interlocutor." McNamara (2008) believes the semi-direct format is reliable for the reason that the interlocutor effect is eliminated. Test takers may also not be nervous as they are not facing a real human but a computer screen, and no interaction is required, but a problem also exists, test-takers are very likely to talk simultaneously in TOEFL speaking test, other participants' voices or even speaking contents may be heard by another test taker, which would also be a distracting factor.

"Provide non-distracting conditions" is also important to ensure reliability (Hughes, 2003, p.48). But there are many other distracting factors influencing test reliability in TOEFL, for example, the test set. Though both TOEFL and ESOL are integrated tests, in TOEFL, speaking is tested as a part of the whole integrated process, even if the speaking test only lasts for 17 minutes, less than it is in ESOL test which is 30 minutes, it cannot be regarded as an independent one. Speaking is the third section in the whole process, the first and second one is the reading and listening test, which have already lasted for a maximum of 130 minutes. Although test administration has offered a 10 minutes break, students may not get fully refreshed from the last tests. Test takers may still be tired and dizzy or even in low spirits if they did not do well in the former two tests (especially the test items in speaking, reading and listening are closely linked, this situation is very likely to happen). While this situation is avoided in ESOL, the speaking is tested independently (though together with listening skill), no former tests that happen within a short time would affect their physical or mental conditions.

The testing condition is also different, in this aspect, TOEFL has lower reliability than ESOL. ESOL is in-person with an examiner and other candidates, and the exam room is quiet, reasonably ventilated and maintained at a comfortable temperature (Trinitycollege, n.d.), these comfortable conditions can facilitate communication, if test-takers feel comfortable and secure, their performance is likely to be as usual, thus the score reliability would increase. TOEFL iBT is a computer-based (students speak into a headset microphone) test, the reliability may be influenced by technical factors, for example, if the microphone has low quality or the headphones produce

noises, these may produce lower test scores.

In reliability, both of ESOL and TOEFL have unreliable factors. With limited aspects concerned, it is hard to decide which one is more reliable.

References:

[1] Alderson, J. C., et al. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

[2] Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

[3] Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

[4] Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language testing*, *20*(1), 1-25.

[5] Clark, H. H. (1979). Responding to indirect speech acts. *Cognitive psychology*, *11*(4), 430-477.Ets TOEFL (n.d.) Retrieved March 20, 2020, from https://www.ets.org/toefl/ibt/about/content/Ets Org (n.d.)Retrieved March 20, 2020, from https://www.ets.org/

[6] Gamaroff, R., (2000). Rater reliability in language assessment: the bug of all bears. *System*, 28(1), pp.31–53.

[7] Hughes, A., (2003). *Testing for language teachers* Second., Cambridge: Cambridge University Press

[8] Lee, J. (2014). *U.S. Patent No. 8,763,098*. Washington, DC: U.S. Patent and Trademark Office.

[9] Luoma, S., (2004). *Assessing speaking*, Cambridge ; New York: Cambridge University Press.

[10] Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. Sage.

[11] Norton, J. (2005). The paired format in the Cambridge Speaking Tests. *ELT journal*, *59*(4), 287-297.

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

[13] O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, *19*(3), 277-295.

[14] Shohamy, E., (1983). THE STABILITY OF ORAL PROFICIENCY ASSESSMENT ON THE ORAL INTERVIEW TESTING PROCEDURES. *Language Learning*, 33(4), pp.527–540.