Journal of Educational Research and Policies ISSN: 2006-1137
DOI: 10.53469/jerp.2024.06(09).26

Comparison and Evaluation of TOEFL iBT and ESOL

Chenxi Yin

Jiangxi University of Technology, Nanchang, Jiangxi, China

Abstract: ***As two high-stake standardized tests, Test of English as a Foreign Language (TOEFL iBT) and Trinity College ESOL Skills for Life (ESOL) Level 2 have been internationally recognized and used to demonstrate candidates' English ability in different skills. Drawing from diverse perspectives, this paper undertakes a comparative evaluation of the speaking sections of these two tests and aims to provide insights into how these exams structure their speaking components, the criteria used for evaluation, and how effectively they measure candidates' English communication skills and practical language abilities. By comparing these aspects, the paper presents the similarities and differences of the speaking assessments of the two testing systems and intends to contribute a deeper understanding within these globally recognized exams to educators, test developers, and candidates alike.***

Keywords: TOEFL iBT, ESOL Level 2, Speaking section, Test construction, Evaluation principles**.**

Tests Comparison: Similarities

In order to analyze a test, it is necessary to have a blueprint about what the test assesses for and how it tests. As the most fundamental concern of all the elements in test construction, the purpose of testing should be firstly taken into account (Alderson, Clapham and Wall, 1995). Every test begins with a specific purpose that relates to what the test is designed for (Fulcher, 2010). With reference to the official website, TOEFL iBT aims to measure candidates' language ability to understand and use English at a certain level. Similarly, ESOL focuses on motivating students to develop and use their communicative and transferable English language skills for everyday life. As such, both tests measure a common aspect of testing the English language ability for a particular purpose, from which can be all regarded as tests of proficiency (Alderson, Clapham and Wall, 1995).

Another common feature is on the skill of ranking that both tests' speaking modules include testing candidates' communicative competence. According to Rivera (1984), communicative competence is a portion of test scores that applied to illustrate language communicative use in a sociocultural context. In the speaking section of TOEFL iBT, candidates' performance will be assessed based on to what extent they can use English to communicate interactively. Similarly, with the objective of assessing language use for everyday life, ESOL speaking tests focus on supporting the development of communicative language skills within a realistic context.

Tests Comparison: Differences

Despite some similarities in types and requisite skill of tests, these two tests are different in more aspects. The major distinction is in the grading policy that they use different methods to assess. The ESOL test is externally assessed by a visiting qualified examiner. In task 2 and 3, the rater plays the role of both examiner and interlocutor. Rather than assessing the candidates, he or she needs to take part in the process and interacts with them (Alderson, Clapham and Wall, 1995). Regarding TOEFL iBT test, the whole test process is administrated via the computer that candidates' performance is captured and digitized automatically. To clarify, there is no interaction between raters and test-takers. In the ranking part, all of the test items will not be rated by the same rater that at least two or three examiners will rate an entire speaking section (Farnsworth, 2013).

Additionally, the TOEFL iBT differs from the ESOL Level 2 in its scoring rubric. When it comes to the TOEFL speaking module, the test descriptors provide four rubric components in detail: general description, delivery, language use and topic development, and each of them will be rated from 0 to 4. The examiners only need to consult two forms of speaking rubrics to grade each task and provide a holistic score in the end (Zahedi and Shamsaee, 2012). However, the examiners in the ESOL are in a different situation that they use the assessment criteria amplification as a reference to measure the performance of the candidates and each task has a different amount of criteria. For example, while test-takers will be assessed from three criteria in task 1 and 2, there are four criteria in task 3 and 4 individually.

What is more, two tests also differ in the task items that involve different types of tasks. When the TOEFL iBT test focuses more on candidates' integrated communicative performance that three of four tasks are integrated, all of the ESOL level 2 speaking test is mainly independent that contain one-to-one conversation and group discussion. Independent speaking constructs are those that measure speaking as a separate skill, in which test-takers speak about a topic without other sources of references like audio and reading texts (Barkaoui et al., 2013). In contrast, the integrated speaking task require test-takers to integrate language skills to understand and incorporate the information, and then transform into spoken response (Frost et al., 2020). The speaking section of TOEFL iBT applies more than one skill that integrated reading and listening into the speaking module. In the integrative part, the candidates should listen to, read, or both some authentic materials in order to provide relative response orally (Zahedi and Shamsaee, 2012). As far as Level 2 of the ESOL speaking section, although the construct itself is intentionally integrated speaking and listening test into one unit, the listening competence has not been required in each task observably. As such, it seems no integrated test tasks are designed in this test.

Tests Evaluation

The overarching principles of test evaluation are validity and reliability that indicate to what extent the test is a good one (Alderson, Clapham and Wall, 1995). As the central concern in evaluating a test, test validation will be highlighted as follow to show whether these two high-stake language tests measure what they are supposed to measure (Weir, 2005). From Messick (1989) perspective, validity is defined as an integrated evaluative judgment based on the evidence of score interpretation and use. Generally speaking, it can be classified into three major types: rational, empirical and construct validity (Alderson, Clapham and Wall, 1995). Above all, construct validation will be primarily used to evaluate these two tests because it is regarded as the most comprehensive part of the validation (Geranpayeh, 2000).

Construct validity involves assessing how well the different test components relate to each other (Alderson, Clapham and Wall, 1995). In the context of speaking portion of the TOEFL iBT, candidates are required to complete three integrated tasks, in which they have to integrate information from reading and listening text before speaking out their opinions (Frost, 2020). Concerning the test content, almost all of the chosen texts and questions deal with social or academic situations, such as a lecture (Farnsworth, 2013). This corresponds to the purpose of the test about improving learners' communicative competence for future studying or career. In the future realistic context, communicative acts depend on the integration of two or more skills that demand test taker's comprehensive language ability (Zahedi and Shamsaee, 2012). In this case, the test seems valid since the task items are correlated well with the purpose of the test.

Furthermore, the TOEFL iBT speaking section applies an automatic format to respond. In this type of format, candidates need to respond with extended monologues into a microphone by using a computer system (Farnsworth, 2013). In this case, examiners will not be a part of the responding process and candidates are required to formulate their answers individually. Based on Hsu and Davidson's (2012) theory of normal face-to-face speaking tests, examiners' instant judgment are more likely to differ from person to person, and potential bias may appear because of their personal characteristics. As a result, judgment about test-takers' performance may be affected, and the test scores may not be accurate evidence to represent their language levels at all (Huges, 2003). On the contrary, without the interruption from human factors in a variety of situations, the test result may equal to every test takers and be more reliable than it from the examiner-student interaction (Weir, 2005). Also, the tape technique is more practical that a large number of candidates can be examined at the same time. As their performances are recorded and will be sent in the form of digital files to be marked, it seems to be a flexible assessing method to the examiners because they can assess candidates' performance at an appropriate time and place (Weir, 2005).

However, things are different in the ESOL speaking test of level 2. For instance, the listening and speaking tests are combined into one section with four independent tasks. Learners are expected to listen and respond to spoken language, which is viewed as an acceptable model in assessing language competence (Frost and Wigglesworth, 2012). What is more, this test provides various types of tasks in the context of a real situation, from which candidates' ability to interact orally may be presented easily. Take group discussion as an example, the examiner takes no part in this student-to-student test that candidates can only communicate with peers to complete it. Tasks of this type cover both interactional and informational routines. Rather than simply repeat rehearsed phrases, candidates are required the ability to produce responses in an unpredicted communicative situation (Weir, 1993). These routines contain both interaction and agenda management skills that test-takers' additional competence like strategic and discoursal competence can be illustrated during the process (Weir, 2005). In this way, their entire interactive performance has a chance to be assessed comprehensively.

As discussed above, due to the reason that the test consists of appropriate tasks with authentic material and matches its purpose, it is fair to say the construct validity of the test has been achieved to some extent. In the perspective of learning outcomes, however, test validity has been diminished on account of test instruction and scoring specification. First of all, it is likely that the test constructors pay more attention to promote candidate' speaking skills such as verbal communication, convey information in the ESOL speaking section. These are mainly involved in the language productive competence, but listening, as a receptive language skill has been overlooked. As illustrated by Frost and Wigglesworth (2012), test construction should be designed to support the descriptor's original intention, which in this test is to assessing listening skills and speaking proficiency in one structure. To achieve that, both speaking and listening components should be included in the context instruction in order to provide a clear framework to the examinees about what they are supposed to achieve at a certain level.

As for the scoring rubric, a controversial situation can be found in the test specifications. On the one hand, the criteria descriptions may be over complicated that raters have to give marks with the reference to two different forms. It is true that instruction should be as clear as possible to enable examiners to follow. Whereas, in order to have a positive effect on the test validity, practicability and flexibility should also be taken into consideration that user specifications need to be accurate and well-constructed (Alderson, Clapham and Wall, 1995). Moreover, when it comes to the specification for candidates, it should be easy to read and measure so that test-takers can find the key strategy of how to improve their language proficiency and what is the standard of their ideal level in s short time (Weir, 2005).

On the other hand, it seems the assessment criteria is not as adequate as it supposed to be. Fulcher (2010) claims that if a test provides information on more than one skill or ability, it should be marked in the form as it is meant to do. ESOL speaking and listening module is a skill-combined assessment that both speaking and listening abilities will be assessed. However, the truth is the generic performance descriptors only include a small portion of listening parts, which highlight the grammatical and phonological features of this competence. As such, examinees' listening ability may not be represented by the holistic scores accurately.

Above all, these two tests provide concrete examples about how good testing practice should be constructed. Although the TOEFL iBT may exhibit certain advantages over the ESOL test in specific aspects, both tests serve as exemplary models of good testing practice, emphasizing the ongoing need for careful consideration and innovation in the field of language assessment. By highlighting the strengths and potential areas for improvement in both tests, it is important to continually refine testing practices to ensure their validity, reliability, and fairness.

References

[1] Alderson, J.C., Clapham, C. & Wall, D., 1995. *Language test construction and evaluation*, Cambridge: Cambridge University Press.

[2] Barkaoui, K. et al., 2013. Test-Takers' Strategic Behaviors in Independent and Integrated Speaking Tasks. *Applied Linguistics*, **34**(3), pp.304–324.

[3] Farnsworth, T.L., 2013. An Investigation Into the Validity of the TOEFL iBT Speaking Test for International Teaching Assistant Certification. *Language Assessment Quarterly*, **10**(3), pp.274–291.

[4] Fulcher, G., 2010. *Practical language testing*, London: Hodder Education.

[5] Frost, K., Elder, C. & Wigglesworth, G., 2012. Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing*, **29**(3), pp.345–369.

[6] Frost, Kellie; Clothier, Josh; Huisman Annemiek; Wigglesworth, Gillian et al., 2020. Responding to a TOEFL iBT integrated speaking task: Mapping task demands and test takers' use of stimulus content. *Language testing*., **37**(1), pp.133–155.

[7] Geranpayeh, A., 2000. *Language proficiency testing: a comparative analysis of IELTS and TOEFL.*

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

[9] Hsu, H.-L. & Davidson, Fred, 2012. *The impact of world Englishes on language assessment: Rater attitude, rating behavior, and challenges*, pp.ProQuest Dissertations and Theses.

[10] Messick, S., 1989. Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher*, **18**(2), pp.5-11.

[11] Rivera, C., 1984. *Communicative competence approaches to language proficiency assessment : research and application*, Clevedon: Multilingual Matters.

[12] Weir, C.J., 1993. *Understanding and developing language tests*, New York ; London: Prentice Hall.

[13] Weir, C.J., 2005. *Language testing and validation : an evidence-based approach*, Basingstoke: Palgrave Macmillan.

[14] Zahedi, K. & Shamsaee, S., 2012. Viability of construct validity of the speaking modules of international language examinations (IELTS vs. TOEFL iBT): evidence from Iranian test-takers. Educational Assessment, Evaluation and Accountability, 24(3), pp.263–277.