Education, Language and Sociology Research
ISSN 2690-3644 (Print) ISSN 2690-3652 (Online)
Vol. 4, No. 2, 2023
www.scholink.org/ojs/index.php/elsr

Original Paper

Development of Scoring Standard of English Oral Test in Respects of Research and Practice

Liu Yixi[1]

[1] Nanjing, China

Received: March 22, 2023 Accepted: April 7, 2023 Online Published: April 18, 2023
doi:10.22158/elsr.v4n2p47 URL: http://dx.doi.org/10.22158/elsr.v4n2p47

Abstract

This paper aims to summarize the development of international oral test scoring standards since 1980s. The development of scoring standards for oral test mainly includes two aspects: research and practice. In the respect of research, the scoring standard of oral test has gone through three stages: "expert experience", "practitioner cognition" and "learner development"; In terms of practice, the scoring standard of oral test can be roughly divided into three types: "native speaker", "being able to express" and "typical characteristics". Based on the analysis of representative research papers and typical practical cases, this paper sorts out the development of international oral test scoring standards in order to provide suggestions for the research and practice of oral test scoring standards in China.

Keywords

oral test, scoring standard, research and practice

1. Introduction

In recent years, many large-scale foreign language tests in China have gradually incorporated English oral tests, such as college entrance examinations in some provinces and cities, CET-4 and CET-6, TEM-4 and TEM-8 and so on (Lv, Song, Wang, Liu, & Li, 2008). Many researchers have actively and effectively explored English oral tests and their research focuses on the design principles and propositions of oral test (Xue, 2009), the use of scoring methods or criteria (Jin, Wang, Song, & Guo, 2008), and the analysis of scoring reliability or rater deviation (Dai, 2010). However, the specific research and development of the scoring standard for oral test is rarely involved.

The function of the scoring standard of oral test is to define the examinee' s oral ability operationally. Based on the scoring criteria, the raters can evaluate the candidates' oral performance and give corresponding scores. With the appearance of oral English test, the scoring standard of oral English test has a long history of development. As early as the 1830s, scoring standards have been developed and applied to oral test, providing relevant data for classroom teaching and school development. Later, because oral English test was widely used in international mobility, higher education, employment and other fields, the scoring standard became the key to provide effective reference. Its development has gradually become one of the focuses of language testing researchers (Fulcher, 2015).

Since the mid-1980s, with the implementation and promotion of ELTS (predecessor of IELTS), IELTS and TOEFL (Taylor & Falvey, 2007), international oral test has made great progress in research and practice, which is typical and has high analytical value. In order to systematically investigate the development of scoring standards for international oral test in the past 30 years, this paper intends to sort out and review the literature and cases from two aspects: research and practice.

In the respect of research, this paper reviews the oral language proficiency scales, test syllabus and significant paper achievements at home and abroad, and sorts out and summarizes the stage characteristics of the development of oral test scoring standards.

In the aspect of practice, this paper analyzes the research and practice results of ELTS, IELTS and TOEFL test organizers or R&D teams, and investigates the transformation of theoretical development related to scoring standards in practical application.

In research aspect, internationally the development of language proficiency standards has gone through three stages: “expert experience”, “practitioner cognition” and “learner performance” (Zhao, Jin, & Wang, 2015). As an important part of the language proficiency standard, the scoring standard of oral test has gone through three important stages as well. They are named as “expert experience”, “practitioner cognition” and “learner development” (Jin & Mak, 2013). On the practical level, the scoring standards of oral test can be roughly divided into three types: “native speaker”, “being able to express” and “typical characteristics” (Jin & Mak, 2013). Its representatives are ELTS 1986 (Davies, 2008), IELTS 1989-2001 (Ingram & Wylie, 1993), IELTS 2001- and TOFEL iBT 2005-. Research and practice have a close relationship and promote each other. However, no matter at the research level or the practice level, the three stages do not suddenly replace the other completely. This paper analyzes the development of scoring standards for oral test in the respects of research and practice.

2. Study on the Scoring Standard of Oral Test

In 1980s, there was a lack of scientific evidence for the formulation of scoring standards. “Expert experience” was often used as the main reference for the development of scoring standards (Fulcher, 2015), that is, language experts took native speakers as benchmarks, compared foreign language learners’ oral English with native speakers, and judged learners’ oral English level according to their similarities. The early Foreign Affairs Association scoring standard was a reference model for many “expert experience” scoring standards. The scoring standard, Interview Assessment Scale, used the concept of score segment, which divided the oral language ability of foreign language learners into nine grades, among which the ninth grade learners reach the level of expert speaker. This level was the highest on and other grades took this as the standard to degrade one by one from the highest level of expert speaker to the lowest level of non-native speaker.

During the 1990s, the grading standard of expert experience gradually exposed the defects and deficiencies of vague grading standards and lack of theoretical and practical basis in the long-term practice (Bachman & Savignon, 1986). In order to develop or revise the scoring standard of oral test, researchers have gradually carried out empirical research based on "practitioners' cognition". Practitioner cognition focuses on how raters, as practitioners, view and evaluate candidates' performance in oral tests. The most typical representative research at this stage is quantitative study of descriptors. Classification Standard Descriptor (North & Schneider, 1998) analyzes the existing scoring standards and combined with teachers' interpretation of standard descriptors. Rasch model is used to quantify the descriptors of learners' oral proficiency standards, which ensures the validity, comprehensibility and mutual independence of the descriptors used, and provides an empirical basis for the development and revision of the scoring standards for oral test.

In the 21st century, both researchers and practitioners argue that scoring standards should take performance of candidates into consideration (Van, 1989). Researchers began to explore learners' language features based on practitioners' cognition, focusing on in-depth analysis of learners' oral ability in different stages of development so as to establish the relationship between oral performance and grade scores. The most typical representative studies at this stage are Brwon's examination of the rating process in the revised IELTS Speaking Test (2006) and study on examination of rater orientations and test taker performance on English-for-academic-purposes speaking tasks (Brown, Iwashita, & McNamata, 2005), but their studies are slightly different. Brown (2006) respectively explored the cognition of IELTS raters and the oral performance of candidates, which were conducted independently. From the perspective of practitioners' description and learners' language features. This paper verifies the validity of the existing IELTS speaking test scoring standards. Brown et al. (2005) combined practitioner description with learners' language features and through a mixed research design, first explored what important features raters paid attention to when evaluating candidates' oral performance. Then, these characteristics are used to analyze the candidates' oral performance, thus verifying the validity of TOEFL oral test scoring standard.

3. Practice of Scoring Standard for Oral Test

In 1980s, under the influence of "expert experience", "native speaker" standard became the main type of scoring standard for oral test in that period. ELTS (Taylor, 2007) was a typical representative in the standard of native speakers at that time. The ELTS scoring standard took native speakers as the reference standard, and divided the oral level of candidates into nine grades. They are expert level, excellent level, good level, qualified level, moderate expression, critical level, extremely limited ability, speaking intermittently, unable to express and absent from examination (Davies, 2008). It can be seen that "expert experience" has an important influence on the development of ELTS scoring standard.

Since 1990s, the study of "Practitioner Cognition" has exerted a certain influence on the practice of grading standards. During the practice of ETLS scoring standard, the raters found its defects and deficiencies when using and interpreting the scoring standard. Therefore, at the end of 1980s, ELTS reviewed and revised its scoring standards (Taylor, 2007). In the revision process, ELTS' scoring standard paid less attention to "expert experience", and its attention turned to "practitioner cognition", that is, it explored how raters evaluate candidates' oral performance (Alderson, 1991), and then develops and revises the scoring standards. In 1989, the revised ELTS scoring standard was put into use and officially changed its name to IELTS (Taylor, 2007).

It can be seen from this that before 1990s, "expert experience" was the mainstream method to construct the scoring standard, and great changes have taken place in the formulation of the scoring standard since 1990s. Different from relying on experts' experience in the past, the rater, the practitioner of oral test, began to play an important role in the process of formulating grading standards. Scoring standards began to pay attention to the rater's cognition of candidates, that was, the rater thinks that the candidates can do what, and then presents them in the way of "being able to express". This kind of "being able to express" based on "practitioners' cognition" had a great influence on the development of the scoring standard of oral test in 1990s, and its representative was IELTS 1989-2001. The grading standards in this edition pay particular attention to what candidates can do in oral performance. For example, in the description of Level 4, it is required that candidates can communicate orally on familiar topics and master basic sentence patterns, grammar and vocabulary (Ingram & Wylie, 1993).

However, the scoring standard of "being able to express" based on practitioners' cognition has gradually exposed its shortcomings: "being able to express" is a concrete description of candidates' oral communicative competence, but it is still vague to a certain extent, and does not really make a detailed analysis and detailed description of candidates' oral performance.

Since the 21st century, the study of "learner development" has greatly influenced the revision and practice of the scoring standard of oral test. The formulation of grading standards further emphasizes the language characteristics of candidates when completing oral tasks. For example, in the IELTS website 2010 standard, the fluency and coherence of Grade 6 candidates can be characterized by occasional repetition, self-correction or hesitation, and the ability to use rich conjunctions; In the TOEFL IBT 2005 standard, candidates with a score of 3 can show the characteristic of individual mistakes in pronunciation.

Compared with the grading standards in 1990s, the grading standards for oral test since the 21st century have added the "typical characteristics" of the candidates' oral performance on the basis of "being able to express", thus making the grading standards develop to a more concrete and more clear which is more convenient for the application of the raters. Practitioner's Cognition introduces the study of scoring standards into empirical research, and begins to pay attention to the raters' attitudes and views on the quality of candidates' oral performance. On this basis, "learner development" analyzes the specific characteristics" of the candidates' oral performance in more detail way. It is precise because of the analysis of these “typical characteristics”. The computer scoring for candidates’ oral presentation is gradually developed.

Computer scoring is an effective combination of second language acquisition, language assessment and computer linguistics. With the continuous improvement of natural language processing technology, computer scoring has developed rapidly (Xi, 2010). In this process, the development of machine scoring standards has become the key to improve the validity of computer scoring. The development of computer scoring standards is mainly based on oral performance characteristics. The researcher further quantifies and digitizes the main and distinctive oral performance features, and realizes the quantitative calculation and automatic scoring of candidates’ oral performance through computer, thus making computer scoring an effective auxiliary and supplement to manual scoring.

4. Discussion

4.1 Change in the Research Respect

Reviewing development of research in the past 30 years, there have been two major changes in the research paradigm. One is the change from “expert experience” to “practitioner cognition” (North & Schneider, 1998), and the other is that based on practitioner cognition, it paid more attention to learner development (Brown et al., 2005). In essence, the change of research paradigm is not accidental. It is determined by the characteristics of the scoring process of oral test itself. In essence, oral test scores are “raters” composed of “experts” and “teachers” who judge the “learner performance” generated by “oral test tasks” based on “scoring criteria” so as to obtain the “score”. In the 1980s, the study of grading standards mainly tried to establish the relationship between “experts” and “scores” through “expert experience”.

However, “expert experience” is not systematic empirical cognition analysis, but experience and intuitive feeling. Therefore, in 1990s, researchers began to look for a more direct empirical basis, based on “practitioners’ cognition”, and directly explored the relationship between “raters” and “scores”. After entering the 21st century, on the basis of “practitioners’ cognition”, researchers have made further explorations. Direct analysis of “learner development”, in order to establish the relationship between “learner development” and “scores”. At the same time, due to the constant change of research paradigm, the research design adopted in the research of grading standards has also changed correspondingly, from the initial state of being lack of empirical research to the single cognitive research and then the study combining cognition and development (Brown, 2006). The research paradigm develops from experience to demonstration, and from being single to being mixed.

4.2 Changes in Practice Respect

Reviewing development of practice in the past 30 years, due to the influence of the research paradigms of “expert experience”, “practitioner cognition” and “learner development”, there have been two major changes in the specific use of scoring standards. One is the change from the standard of “native speaker” to the standard of “being able to express”. The other is the change from the standard of “being able to express" to more detailed standard of "typical characteristics" (Chapelle, 2012). In essence, these two changes in scoring standards are also different interpretations of the construction of spoken language ability by language test practitioners. The standard of "native speaker" emphasizes the similarity of native speaker's language imitation. The criterion of "being able to express" and the criterion of "typical characteristics" emphasize the learners' communicative competences. The difference between the two is that the former pays attention to what can be done, while the latter pays attention to the specific characteristics that learners show in the process of "what can be done". It can be seen that the practice development in the past 30 years has developed from the general definition of "native speaker", the specific definition of "being able to express" to detailed description of "typical features". The scoring standard of oral test is changing from being general to being concrete and from fuzzy to precise (Jin, Mak, & Zhou, 2012). The grading standard of oral test is changing and developing towards concretization and precision. This will provide raters with more reliable scoring basis and learners with more effective oral test scores, thus further improving the validity and reliability of oral test practice.

4.3 Development Trends and Enlightenment

Since the 21st century, both researchers and practitioners are more inclined to adopt the mixed research paradigm (Creswell, 2013) to explore the development of "typical characteristics" standard (Brown et al., 2005). It can be predicted that in the future development of scoring standards, the "characteristics" of learners' specific language use and performance will become the focus of research and practice. Two hot research topics have emerged. The first topic focuses on the selection and use of "features". Because of the complexity of speaking ability itself, how to choose representative and distinctive features to construct scoring standards (Jin & Mak, 2013). Whether the same features should be used in every grade of the scoring standard (Humphry & Heldsinger, 2014) will become a hot issue in future research. The second topic focuses on computer scoring based on "features". Because some "features" can be quantified, selecting more important and simplified "features" for quantification, and performing machine scoring by computer can assist manual scoring. In the future, we can explore how to improve the accuracy and reliability of machine scoring, so that machine scoring can more effectively assist and cooperate with manual scoring, thus further improving the scoring quality of oral test. However, it should be pointed out that whether it is machine scoring or manual scoring, both of them will face more challenges on the theoretical level, such as the evaluation methods of interactive ability, the evaluation methods of discourse co-construction and the evaluation criteria of comprehensive test tasks.

The development of scoring standards for international oral tests has brought inspiration to the development of large-scale oral tests in China. The typical representative of large-scale oral test in China is CET-4 and CET-6 since 1999, after 20 years of efforts and attempts, a set of nearly perfect oral examination system has been gradually established (Jin & Guo, 2002). From 2016, the national college oral English test will be divided into CET-4 and CET-6. Level 4 is composed of self-introduction, short reading, short answer, personal statement and group interaction. Level 6 consists of self-introduction and question and answer, speech and discussion and question and answer. These two oral examinations can comprehensively assess and evaluate students' oral English expression ability. According to different oral test tasks, China can also build corresponding oral sample database for candidates in the future. The corpus analysis based on the oral sample database can explore the important "characteristics" that affect the oral level of Chinese college students, so as to develop the scoring standard of oral test based on "learners' language features". Furthermore, on the basis of discovering "learners' language features", we can try to quantify candidates' oral performance by using machines and computers in the future to promote the implementation of machine-assisted manual scoring. In the scoring result report of the future oral test, in addition to reporting the score level, it can also provide corresponding "feature" descriptions of candidates of different grades, and provide suggestions for the improvement of candidates' oral ability.

5. Conclusion

The development of scoring standards for oral test in the past 30 years has witnessed the transformation of research paradigms of "expert experience", "practitioner cognition" and "learner development", and also contributed to the development of scoring standards from "native speaker" standard, "being able to express" standard to "typical characteristic" standard. In these years, the improvement of researchers' understanding of language ability and the deep understanding of test construction promote the development of oral test scoring standards at the theoretical level; The enrichment of oral test methods by practitioners, the development of discourse analysis methods and the improvement of oral performance analysis methods of candidates have promoted the development of scoring standards of oral test in practice.

It should be pointed out that there is no absolute boundary between the research paradigms of the three stages. To a certain extent, the description and discussion of language competence framework by "expert experience" guides the research of "practitioner cognition" and "learner development". To a great extent, "practitioner cognition" is the first research and empirical basis of "learner development". The same is true in practice level. Although the standard of "native speaker" has gradually retired from the historical stage, its influence still exists. At present, the standard of "being able to express" plays an important role in the world, and the standard of "typical characteristics" is only in its infancy, which needs further research and practice revision.

The change of research paradigms of "expert experience", "practitioner cognition" and "learner development" has also prompted test researchers to critically apply the relevant research results of linguistics, especially the standard and measurement of language features themselves. After implementing the research paradigm of "learner development", test researchers should pay more attention to the study of L2 learners' own language features, instead of taking native speakers' oral features as a successful paradigm and comparing them with L2 learners' oral features to judge whether they are good or bad (Cook, 2016). At present, the quantitative research on language itself has developed from the surface word feature calculation to the deep syntactic and semantic calculation (McNamara, Graesser, McCarthy, & Cai, 2014). However whether the quantitative calculation and evaluation methods of spoken language features of mother tongue or first language are fully applicable to second language learners needs to be further explored.

In China, with the gradual acceptance of oral English tests in large-scale foreign language tests, the research and practice of oral English test scores have made some achievements. However, empirical studies on the development of scoring standards are rare. In the practice of grading standards for oral test in China, the standard of "being able to express" is basically adopted, and the standard of "typical characteristics" is only partially adopted. In the future research and practice, we can explore the dimensional composition of Chinese learners' oral ability, analyze the important "characteristics" that can distinguish learners' level, and further develop the "typical characteristics" standard, so as to improve the scoring standard of Chinese oral test. In the future, China can also consider exploring the research and application of oral machine scoring in large-scale oral test scoring.as an effective aid to manual scoring.

References

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson, & B. North (eds.). *Language Testing in the 1990s: The Communicative Legacy* (pp. 71-86). London: Macmillan Publishers.

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL Oral Interview. *Modern Language Journal*, *70*(4), 380-390. https://doi.org/10.1111/j.1540-4781.1986.tb05294.x

Brown, A. (2006). An examination of the rating process in the revised IELTS Speaking Test. In P. McGovern, & S. Walsh (Eds.). *IELTS Research Reports 2006* (pp. 41-701). Canberra & Manchester: IELTS Australia and British Council.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An Examination of Rater Orientations and Test Taker Performance on English-for-Academic-Purposes Speaking Tasks (TOEFL Monograph No. 29)*. Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2005.tb01982.x

Chapelle. (2012). Validity argument for language assessment: The framework is simple. *Language Testing*, *29*(1), 19-27. https://doi.org/10.1177/0265532211417211

Cook, V. (2016). Where is the native speaker now? *TESOL Quarterly*, *50*(1), 186-189. https://doi.org/10.1002/tesq.286

Creswell, J. W. (2013). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Thousand Oaks: Sage.

Dai, Z. H., & You, Q. D. (2010). An analysis of the Grader Bias in College English Oral Computer Test. *Foreign Language World*, *5*, 87-95.

Davies, A. (2008). *Assessing Academic English*. Cambridge: Cambridge University Press.

Fulcher, G. (2015). Assessing second language speaking. *Language Teaching*, *48*(2), 198-216. https://doi.org/10.1017/S0261444814000391

Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher*, *43*(5), 253-263. https://doi.org/10.3102/0013189X14542154

Ingram, D. E., & Wylie, E. (1993). Assessing speaking proficiency in the International English Language Testing System. In D. Douglas, & C. Chapelle (Eds.), *A New Decade of Language Testing Research: Selected Papers from the 1990 Language Testing Research Colloquium* (pp. 220-234). Alexandria, VA: TESOL,

Jin, T., & Mak, B. (2013). Distinguishing features in scoring L2 Chinese speaking performance: How do they work? *Language Testing*, *30*(1), 23-47. https://doi.org/10.1177/0265532212442637

Jin, T., Mak, B., & Zhou, P. (2012). Confidence scoring of speaking performance: How does fuzziness become exact? *Language Testing*, *29*(1), 43-65. https://doi.org/10.1177/0265532211404383

Jin, Y., & Guo, J. K. (2002). A study on the Validity of non-interview oral test of College English Test Band 4 and Band 6. *Foreign Language World*, *5*, 72-79.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, *15*(15), 217-263. https://doi.org/10.1177/026553229801500204

Taylor, L. (2007). Introduction. In L. Taylor, & P. Falvey (Eds.). *IELTS Collected Papers: Research in Speaking and Writing Assessment* (pp. 1-34). Cambridge: Cambridge University Press.

Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly*, *23*(3), 489-508. https://doi.org/10.2307/3586922

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, *27*(3), 291-300. https://doi.org/10.1177/0265532210364643

Xue, R. (2009). On communicative oral test and its scoring method. *Foreign Language Education*, *6*, 62-66.

Zhao, W., Jin, T., & Wang, B. R. (2015). The development of college English Language competence standards: Theory, Practice and Enlightenment. *Modern Foreign Languages*, *1*,102-111+147.