International Journal of Social Science And Human Research

ISSN(print): 2644-0679, ISSN(online): 2644-0695

Volume 06 Issue 01 January 2023

DOI: 10.47191/ijsshr/v6-i1-04, Impact factor- 5.871

Page No: 22-30

Computer-Based Vs. Face-to-Face Speaking Assessment: Fitness for Purpose from a Communicative Language Testing View

Yazan Brahim
Sultan Qaboos University, Oman

ABSTRACT: Thanks to the ever-developing educational technologies, more and more educational institutions and testing organizations around the world have been delivering their testing events through computer-based formats. As far as language assessment is concerned, these developments have led to questions on all sorts of validity being raised by educational and assessment researchers, especially regarding the speaking skill. This exploratory paper compares and contrasts the pros and cons of computer-based and face-to-face assessment of speaking from a communicative view of language, by exploring what exactly needs to be assessed in speaking, the effects of the delivery mode and the differences between computerized and human rating. The findings are that face-to-face assessment of speaking is a much more valid format than computer-based assessment from a communicative point of view, but that the latter can provide a partial answer to the bulk administration problem in contexts such as placement or exit tests in educational institutions.

KEYWORDS: computer-based assessment, face-to-face assessment, communicative testing, communicative language ability, interactional ability

1. INTRODUCTION

Language assessment has traditionally taken the form of testing knowledge about language (Brown, 2003). Until the early 80's, this was reflected in discrete-point tests, mainly focusing on vocabulary and grammar, but also on the listening and reading skills, and, to a lesser extent, on the writing skill. Yet as far as the speaking skill is concerned, many international proficiency tests did not include a speaking component, and this continues to be the case of many institutional placement tests around the world. The main reason behind this is that speaking testing events are deemed to be time consuming and difficult to administer on a large scale (Luoma, 2004). The emergence of the *Communicative Language Approach* in foreign language teaching in the late 70's and early 80's, however, constituted a turning point, resulting in a set of fundamental changes in perceptions about what *linguistic ability* is (Canale & Swain, 1980; Bachman & Palmer, 1990). The focus has shifted from *what learners know* about the language to *what they can do* with the language. Such perceptions, in turn, entailed a reconsideration of what needs to be tested, giving birth to the notion of *Communicative Language Testing* of foreign language learners (CLT), and inducing more and more researchers to advocate that communicative syllabus design and communicative methodology be matched by what has been termed 'communicative testing' (Carroll, 1983; Wesche, 1983; Weir, 1988; Katsumasa, 1997). Speaking, as the primary means of communication, has now gained prime importance in both teaching and testing. And because assessing speaking, by its very nature, does not easily lend itself to bulk administration, some international English language proficiency tests have created their own computerized speaking test versions (e.g. Test of English as a Foreign Language (TOEFL) and Pearson Test of English (PTE)), with the intent to make it more readily accessible to examinees around the world.

This latter development, however, has brought about much heated debate among linguists as to whether computer-based testing of speaking (CBT), be it direct or semi-direct, can offer a valid equivalent to direct, face-to-face speaking assessment. This paper will address this question from a communicative language testing point of view by focusing on three main aspects of the speaking assessment: the definition of what constitutes the communicative speaking test construct, the relevant task types and assessment criteria that fit in CBT and face-to-face speaking tests, and the differences between human and computerized rating.

2. DEFINING COMMUNICATIVE SPEAKING TEST CONSTRUCT

Decisions on what to test should reflect what we take language to be, its very nature and the complex underlying principles that govern its usage and use. Larsen-Freeman and Cameron (2008) state that language shows many, if not all, of the characteristics of a complex, dynamic system, involving the interaction of many components, namely linguistic and non-linguistic, verbal and non-

verbal, to convey information, emotion and attitude. Therefore, any good communicative speaking test should always strike a balance between measuring learners' knowledge of the language and *how* they use it. In other words, a good test, from a communicative point of view, should be one that can elicit enough language samples to fully account for testees' communicative language ability, which encompasses two distinct competencies: language competence and strategic competence.

2.1 Language Competence

Bachman and Palmer (1996) built the best-known language assessment model in the field, the Communicative Language Ability model (CLA), which they structured after one that was produced earlier by Canale and Swain (1980). In this model, language Competence is made up of three types of competencies: grammatical, discourse and pragmatic. Grammatical competence has to do with the control and accuracy of syntax, morphology, vocabulary and pronunciation. Discourse competence is the way a speaker plans, organizes and produces his talk through rhetorical organization, coherence and cohesion. Pragmatic competence refers to the speakers' and listeners' ability to deal with the openness of meaning, conversation implicature and talk in interaction. All three competencies contribute to the speaker's level of fluency, which can be defined as the ability to produce rapidly flowing natural speech. Thus, a good speaking test should aim at assessing speakers' language competence as it should reflect their mastery and control over the mechanisms of the language. Research has proved that CBT does as equally well as face-to-face tests in assessing speakers' language competence (Brown, 1993; Luoma, 1997; Mousavi, 2007; O'Loughlin, 2001; Shohamy, 1994). However, in order to capture the whole stretch of speakers' communicative language abilities, we also need to assess their strategic competence.

2.2 Strategic Competence

According to the same CLA model by Bachman and Palmer, strategic competence is made up of two competencies: interaction skills and non-verbal features of interaction. Interaction skills refer to features like flexibility, adaptability and appropriacy of produced utterances, depending on such variables as context, audience or interlocutor. Interaction skills are part and parcel of our overall communicative competence, in that speaking, more often than not, takes place as an oral interaction between two or more people. And it is true that "conversations with different people turn out to be different…because speakers react to each other and construct discussions together" (Luoma 2004, p. 27). In a nutshell, interaction in communication refers to the fact that "participants adopt various devices of conversation according to specific interactional contexts involving interlocutors" (Iwashita, 2021, p. 70). The second pillar of strategic competence is the use of non-verbal means of communication. Meaning in the spoken language, more than in the written one, cannot be fully accounted for without interpretation of the non-linguistic features that always accompany utterances. Indeed, in real life, with the exception of telephone conversations, no spoken utterance is ever free from such factors as gestures, body posture, eye contact or facial expressions, which sometimes speak louder than words.

With the advent of CLT, more and more focus has been placed on the assessment of interaction ability (IC), a concept first introduced by Kramsch (1986), and many commercial tests have started incorporating IC features in their speaking tests. As will follow through this paper, the inability to fully assess speakers' strategic competence constitutes one of the major weaknesses of CBT. Indeed, the unavailability of an interlocutor not only means absence of interaction, but also, as will be shown next, has a direct impact on the testee's performance and output. Logically enough, only a human assessor can detect and interpret the power of the above-discussed non-verbal interaction features.

3. COMPUTER-BASED Vs. FACE-TO-FACE SPEAKING ASSESSMENT: RELEVANT TASK TYPES AND ASSESSMENT CRITERIA

3.1 Benefits and limitations of face-to-face assessment

All the advantages of face-to-face speaking assessment revolve around the very presence of an interlocutor and/or assessor. For one thing, the fact that there is an interlocutor ensures that interaction takes place, that there is two-way communication as there would be in real life. And since interaction is sure to take place, test designers can and should include interaction skills as one of the assessment criteria (see appendix 1 where 'interaction' is one of the Common European Frame of Reference (CEF) oral assessment criteria). This is because speakers' interaction skills, whether verbal or non-verbal, are part and parcel of their overall strategic competence, and therefore should be assessed as equally as their language competence. This leads to the issue of authenticity in the sense of the possibility to design real-life tasks. Linguists have differentiated various types of speaking tasks, probably best summarized by Bygate (1987) as *factually oriented talk* (description, narration, instruction and comparison) and *evaluative talk* (explanation, justification, prediction and decision). If in real life factually oriented talk can happen in the monologic mode, as in storytelling, lectures or presentations, evaluative talk, on the other hand, lends itself more naturally to the two-way interactive, dialogic mode for best production results. This further emphasizes the importance of the interlocutor whose role is to stretch the testees' speaking performance by eliciting more language samples through scripted and/or supplementary questions, as well as to help candidates re-engage in the conversation when they 'dry out' through prompts and cues (see appendix 2 for more information on the role of the examiner in the International English Language Test of English (IELTS) speaking test).

On the negative side, however, many studies have spotted at least two problems with face-to face assessment of speaking: *rater reliability* and the *examiner effect*. The rater reliability problem can arise from all sorts of bias that an examiner can have against or in favour of a testee, or from purely human factors like lack of focus, fatigue or marking by impression rather than by applying specific assessment criteria. The rater reliability problem is usually addressed through good training and pre-test standardization. The second problem, examiner/interlocutor effect, arises from the fact that 'the interviewer has considerable power over the examinee' (Luoma, 2004, p. 35) as it is the interlocutor who initiates and controls the interaction. Other causes of the examiner effect may include personality (Berry, 2007; Van Moere, 2006;), proficiency level (Iwashita, 1996; Nakatsuhara, 2006, 2011) and gender (Brown & McNamara, 2004; O'Sullivan, 2000).

While the possible negative impacts of the above-discussed shortcomings in face-to-face assessment of speaking are valid and genuine, the fact remains that its benefits by far outweigh its disadvantages from a communicative view of language. As will be explained in the next part, the problems associated with the computer-based assessment of speaking are much more compromising than those of face-to-face assessment.

3.2 Benefits and Limitations of Computer-based Testing:

The most controversial issue with computer-based assessment of speaking is probably that of construct validity. Kiddle & Kormos point out that "a number of issues underlie the reservations and generally cautious approach to the use of computer technology in the assessment of speaking…first and foremost… the threat of construct underrepresentation through the lack of interaction in computer-based tests" (2011, p. 342). This means that the very nature of the mode of response in CBT does not cater for the elicitation or accurate assessment of examinees' interactive skills and discourse-management aspects of overall proficiency. As explained above, it is almost impossible to include genuine interactive, dialogic tasks in CBT. Evaluative talk tasks, which ideally require two-way interaction for better spoken production results, are rarely used in CBT. If any, they will take the form of a monologic long turn, which is better suited for assessing presentation skills rather than interaction skills. Therefore, the tasks that are most widely used in CBT assessment of speaking are structured speaking tasks to assess all but interaction skills. Examples of these include reading aloud and sentence repetition (processing-oriented) and sentence completion and factual short-answer questions (grammatical knowledge and contextual understanding); for longer stretches of speech, CBT offers tasks like reacting to phrases or situations and giving a presentation (see appendix 3 for examples of structured task types in a PTE Academic speaking test). It is very difficult to imagine how such tasks can assess the highly unpredictable and creative elements of speaking. And although PTE tries to include tasks that are intended to tap into testees' interactional competence, e.g., presenting them with situations and asking them to respond, Plough et al maintain that "proactive strategies (such as checking comprehension) that occur naturally in communication cannot be operationalized in scripted prompts. This means that tightly scripted tasks risk narrowing the focus of the IC evaluation to the types of IC that can be elicited… [therefore] they do not truly capture IC." (2018, p. 431). In addition, while face-to-face speaking tests may use such structured tasks like the ones used by CBT, the major difference resides in the fact that the latter cannot make use of tasks that require dialogic spontaneity, simply because a machine cannot interact like a human being. Clark probably best echoes this idea: "[in semi-direct speaking tests] the interactive discourse-management aspects of the student's overall speaking proficiency cannot be readily elicited (or by the same token, effectively measured)" (1986, p. 4). In this sense, the construct validity problem with CBT can be explained by the fact that CBT is more concerned with the spoken production than with the spoken interaction which happens in real life, whereas, ideally, test performance and scores should help predict examinees' ability to cope with non-test situations. The direct implication of this construct validity problem is that interaction skills are not part of CBT's assessment criteria, thus missing out on a very important aspect of overall communicative competence (see appendices 3 and 4 where interaction is not part of the assessment criteria in PTE Academic and TOEFL tests). Even the IELTS speaking assessment criteria do not measure interaction ability, which is a big disadvantage, but they do account for the difference between natural and unnatural hesitation (see the task response criterion, appendix 5). In addition, much to the credit of the IELTS exam, even when they have recently launched their new computer-delivered test version, they are only using it for the listening, reading and writing skills, and are still opting for the face-to-face speaking assessment format as it feels more natural and life-like, encourages more and better language output, and, as will be demonstrated in what follows, eliminates problems associated with the mode of delivery.

A no less problematic area in CBT speaking assessment is the non-authenticity of the mode of interaction. Speaking in CBT is one-directional, where the testee is supposed to accommodate to the computer, whereas the computer cannot accommodate to the testee. This in turn disadvantages the examinees in at least two obvious ways: absence of the *communication repair* option and the so-called *delivery mode effect* or *Test Method Effect* (Bachman & Palmer, 1996). The communication repair option, which is categorized by the Common European Frame of Reference as one of the six microfunctions of language use, refers to the possibility for the listener and/or speaker to signal non-understanding and to ask for assistance or rewording (Luoma, 2004, pp. 33-34). This is something that happens all the time in real life, but which testees cannot obviously have recourse to in a CBT speaking test. As for the test method effect, it has to do with "the degree of fairness, fitness for purpose, enjoyment, confidence and comfort with an

electronically-delivered test via hardware and online platforms" (Kiddle and Kormos, 2011, p. 345). For example, can the degree of familiarity (or unfamiliarity) with the keyboard, screen and microphone affect the examinees' performance and therefore affect their scores? If yes, then the test delivery mode is said to cause *construct-irrelevant variance*. Not much conclusive research has been conducted on this particular area, but the available evidence seems to suggest that most test takers consider face-to-face assessment as a fairer way to test their speaking ability (Dean, 2008; Luoma, 1997; Qian, 2009). For instance, in a study that compared IELTS speaking test takers in two different modes, Nikatsuhara et al report that "72% of test-takers and 50% of examiners preferred the face-to-face mode [to the video conferencing mode]" (2017, p. 57). In a similar vein, in an action research to compare the test method effect in CBT and face-to-face testing of speaking, Kiddle & Kormos (2011) found out that the correlation between scores in both modes are generally high, but that candidates scored significantly less in pronunciation in CBT, but scored more in task achievement in the same mode. The lower pronunciation scores in CBT can be explained either by the so-called microphone anxiety, or by technical problems that can interfere with the delivery of the test. Testee anxiety in CBT is also generally thought to originate from the inability to rely on paralinguistic channels, like gestures or facial expressions, to support one's speaking performance. A very revealing conclusion by Luoma (2004, p. 45) suggests that CBT and face-to-face speaking assessments are different as discourse events, and that examinees' language in CBT tends to be 'more literate and less oral-like'. This is a clear indication that examinees do not talk to a human being in the same way they talk mono-directionally to a machine.

Computer-based assessment of speaking, on the other hand, has some advantages. The most obvious benefit is the possibility of bulk administration (Luoma, 2004). Thanks to the huge advancement in computer technology, CBT can be taken simultaneously by thousands of people around the world. Here the main concern is not with the potential of huge profits that some international proficiency tests like TOEFL and PTE are making, but rather with the possibility of offering the test to hundreds of students at school or university level, either as placement or exit tests. This will help place people at the right level for better learning results. And despite all the above-mentioned limitations of CBT, offering it in such contexts remains better than doing away with the speaking test altogether. Another benefit of CBT is scoring reliability (Luoma, 2004). This has to do with the elimination of the human bias factor, as a computer will score all candidates in the same way.

4. HUMAN OR COMPUTERIZED RATING?

Whether computers can assess spoken communication as accurately as human raters remains a big question. Obviously, all the above discussion about the complexity of the communicative language ability and interaction ability largely pleads in favour of human rating. In the case of objectively ratable tasks, it could very well be argued that CBT can assess spoken production as accurately as human raters. Streeter et al (2011), for instance, claim that PTE test designers have been able to prove this through highly-documented extensive research, pre-testing, trialing and validation work (see appendix 6 for machine-human and human-human score correlations). However, many linguists remain skeptical as to the ability of CBT to fully detect and accurately assess all complex aspects of pragmatic as well as strategic competence. Meaning, in particular, is probably the most elusive aspect of language. The way Elliott (2010, p. 16) puts it, 'In reality, the production of meaning is a highly complex process involving the interaction of a variety of components: lexis, grammar, phonology, discourse-level features, paralinguistic and non-verbal features and, crucially, context.' Therefore, how can a machine capture the non-explicitness or the open nature of meaning, which are very advanced and effective strategies in productive skills, and where a speaker might say something to mean another? How can a machine account for the non-linearity of spoken language, where "a slight change in intonation could render a completely different interpretation to an utterance" (Elliott, 2010, p. 16)? How does a machine differentiate unnatural from natural hesitation, the latter being a characteristic of a person's talk, and which can be due to planning for argument rather than fumbling with words? Can a computer fully recognize the functions of small words in a conversation, like "well" or "you know", these 'discourse lubricants' which Luoma considers as "a marker of highly advanced speaking skills" (2004, pp. 16-17)? As a matter of fact, no matter how advanced the technology is, these are language aspects which are so peculiarly human that only human raters can account for. Indeed, the productive communication skills needed, at least in spoken mode, are ones that contain a strong component of non-verbal features, as well aspects of discourse management that are not objectively scorable.

5. CONCLUSION

In summary, a good speaking assessment, from a communicative language testing point of view, should be one that can accurately evaluate and objectively score speakers' overall communicative language ability. In order to achieve this, test designers should aim at assessing both the language competence (grammatical, discourse and pragmatic) as well as the strategic competence which includes interaction skills and non-verbal communication aspects. Tasks that best suit this purpose should be presented in both the monologic and the dialogic modes, using relevant topics to elicit factual talk as well as evaluative talk. It logically follows that the assessment criteria should include interaction skills as one of the competencies to be measured. For scoring, well-trained human assessors should be used and pre-test standardization should be conducted to minimize the subjectivity factor that can be associated with human rating. Taking all the above into consideration, face-to face tests are a much more valid, natural and life-like format of

assessing speaking than CBT, as it is much better equipped to tap in the testees' communicative competence and interaction ability. And despite the fact that some researches have established high correlations between scores from both test modes, the fact remains that there is an obvious construct underrepresentation in CBT, in the sense that it is missing out on many of the basic properties of language as a complex, multidimensional system.

As for CBT assessment of speaking, it can present a partial answer to the problem of bulk administration of speaking tests in schools and universities in the context of placement and/or exit tests.

REFERENCE

1) Bachman, L. F. & Palmer, A. S. (1996). *Language testing in Practice*, Oxford, UK: Oxford University Press
2) Brown, A. (1993). The role of test-taker feedback in the test development process: test takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing*, 10, 277–304
3) Brown, A., & McNamara, T. (2004). "The devil is in the detail": Researching gender issues in language assessment. *TESOL Quarterly*, 38, 524–538.
4) Brown, H. D. (2003). Language assessment: Principles and classroom practices. White Plains, NY: Longman.
5) Bygate, M. (1987). *Speaking*. Oxford: OUP.
6) Canale, M. & Swain, M. (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics*, 1 (1), 1-48
7) Clark, J. L. D. (1986). *Handbook for the development of tape-mediated ACTFL/ILR scale-based tests of speaking proficiency in the less commonly taught languages*, Washington, DC: Centre for Applied Linguistics. [Google Scholar]
8) Carroll, B. J. (1983). Communicative language tests: Tasks, enabling skills, formats, and measurement criteria. World Language English, 2(1), 37-39.
9) Dean, M. (2008). *Man or machine: Korean stakeholders' views on oral proficiency assessment (Unpublished master's dissertation)*, Lancaster, United Kingdom: Lancaster University. [Google Scholar]
10) Elliott, M. (2010). The expression of affect in spoken English, *Cambridge ESOL: Research Notes*, 42 (4),16-21
11) Iwashita, N. (1996). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing*, 5(2), 51–65.
12) Iwashita, N. (2021). UNDERSTANDING THE CONSTRUCT OF SPEAKING PROFICIENCY AND ITS IMPLICATIONS FOR CLASSROOM-BASED ASSESSMENT. *European Journal of Applied Linguistics and TEFL*, 10(1), 63-78. Retrieved from https://search.proquest.com/scholarly-journals/understanding-construct-speaking-proficiency/docview/2516295372/se-2?accountid=27575
13) Katsumasa, S. (1997). Communicative language testing: Principles and problems. English Review, 12, 3-24.
14) Kiddle, T. & Kormos, J. (2011). The Effect of Mode of Response on a Semidirect Test of Oral Proficiency, *Language assessment Quarterly*, 8 (4), 342-360
15) Kramsch, C. (1986). From language proficiency to interactional competence. The Modern Language Journal, 70(4), 366–372.
16) Larsen-Freeman, D. & L. Cameron. (2008). *Complex systems and applied linguistics.* Oxford: Oxford University Press.
17) Luoma, S. (1997). *Comparability of a tape-mediated and a face-to-face test of speaking: A triangulation study* (Unpublished licenciate thesis). Centre for Applied Language Studies, University of Jyvaskyla, Jyvaskyla, Finland https://jyx.jyu.fi/dspace/bitstream/handle/123456789/11733/337.pdf?sequence=1
18) Luoma, S. (2004). *Assessing Speaking*, Cambridge: Cambridge University Press
19) Nakatsuhara, F. (2006). The impact of proficiency level on conversational styles in paired speaking tests. *Cambridge ESOL Research Notes*, 25, 15-19
20) Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language testing*, 28 (4), 483-508
21) Nakatsuhara, F., Inoue, C., Berry, V. and Galaczi, E. (2017). Exploring performance across two delivery modes for the IELTS Speaking Test: face-to-face and video-conferencing delivery (Phase 2). *IELTS Partnership Research Papers*, 3, 1-74. Available at https://www.ielts.org/teaching-and-research/research-reports
22) Mousavi, S. A. (2007). *Development and validation of a multimedia computer package for the assessment of oral proficiency of adult ESL learners: Implications for score comparability (Unpublished doctoral dissertation)*, Australia, Queensland: Griffith University. [Google Scholar]
23) Plough, I., Banerjee. J, & Iwashita, N. (2018). Interactional competence: Genie out of the bottle. *Language Testing*. 35(3), 427-445.
24) O'Loughlin, K. (2001). *Studies in Language Testing: The equivalence of direct and semi-direct speaking tests*, Cambridge, UK: Cambridge University Press

25) O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System*, 28, 373–386
26) Qian, D. (2009). Comparing direct and semi-direct modes for speaking assessment: affective effects on test takers. *Language Assessment Quarterly*, 6, 113–125. [Taylor & Francis Online], [Web of Science ®], [Google Scholar]
27) Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing*, 11, 99–123
28) Streeter, L., Bernstein, J., Foltz, P., & DeLand, D. (2011). Pearson's Automated Scoring of Writing, Speaking, and Mathematics, *Pearson White Papers [Online]*, June 2011, available at: https://images.pearsonassessments.com/images/tmrs/PearsonsAutomatedScoringofWritingSpeakingandMathematics.pdf
29) Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing*, 23(4), 411-440.
30) Weir, C. J. (1988). Communicative language testing with special reference to English as a foreign language. Exeter Linguistic Studies, 11, 1-241.
31) Wesche, M. B. (1983). Communicative testing in a second language. The Modern Language Journal, 67(1), 41-55.

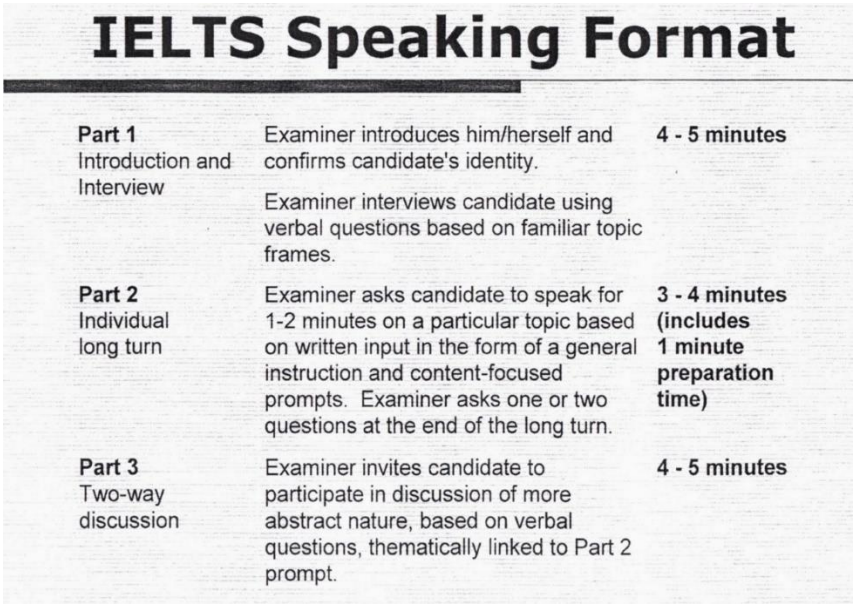
Appendix 1. CEF Oral Assessment Criteria. Source: Cambridge ESOL Examinations

UNIVERSITY of CAMBRIDGE
ESOL Examinations
English for Speakers of Other Languages

Table 5.5: ORAL ASSESSMENT CRITERIA GRID (CEF Table 3)

	RANGE	ACCURACY	FLUENCY	INTERACTION	COHERENCE
C2	Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).	Can express him/herself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it.	Can interact with ease and skill, picking up and using non-verbal and intonational cues apparently effortlessly. Can interweave his/her contribution into the joint discourse with fully natural turntaking, referencing, allusion making etc.	Can create coherent and cohesive discourse making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices.
C1+					
C1	Has a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say.	Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur.	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.	Can select a suitable phrase from a readily available range of discourse functions to preface his remarks in order to get or to keep the floor and to relate his/her own contributions skilfully to those of other speakers.	Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices.
B2+					
B2	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some complex sentence forms to do so.	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, and can correct most of his/her mistakes.	Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he or she searches for patterns and expressions, there are few noticeably long pauses.	Can initiate discourse, take his/her turn when appropriate and end conversation when he/she needs to, though he/she may not always do this elegantly. Can help the discussion along on familiar ground confirming comprehension, inviting others in, etc.	Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some "jumpiness" in a long contribution.
B1+					
B1	Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.	Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable situations.	Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.	Can initiate, maintain and close simple face-to-face conversation on topics that are familiar or of personal interest. Can repeat back part of what someone has said to confirm mutual understanding.	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.
A2+					
A2	Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations.	Uses some simple structures correctly, but still systematically makes basic mistakes.	Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident.	Can ask and answer questions and respond to simple statements. Can indicate when he/she is following but is rarely able to understand enough to keep conversation going of his/her own accord.	Can link groups of words with simple connections like "and", "but" and "because".
A1+					
A1	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.	Shows only limited control of a few simple grammatical structures and sentence patterns in a memorised repertoire.	Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.	Can ask and answer questions about personal details. Can interact in a simple way but communication is totally dependent on repetition, rephrasing and repair.	Can link words or groups of words with very basic linear connectors like "and" or "then".
Below A1					

Appendix 2. IELTS Speaking Test Format. Source: Cambridge ESOL Examinations



IELTS Speaking Format

Part 1 Introduction and Interview	Examiner introduces him/herself and confirms candidate's identity. Examiner interviews candidate using verbal questions based on familiar topic frames.	**4 - 5 minutes**
Part 2 Individual long turn	Examiner asks candidate to speak for 1-2 minutes on a particular topic based on written input in the form of a general instruction and content-focused prompts. Examiner asks one or two questions at the end of the long turn.	**3 - 4 minutes (includes 1 minute preparation time)**
Part 3 Two-way discussion	Examiner invites candidate to participate in discussion of more abstract nature, based on verbal questions, thematically linked to Part 2 prompt.	**4 - 5 minutes**

Appendix 3. PTE Item Scoring (part 1). Source: PTE Academic Score Guide
https://pearson.com.cn/file/PTEA_Score_Guide.pdf

Part 1 Speaking and Writing (approx 77–93 minutes)

Item type	Time allowed	Number of items	Scoring	Communicative skills, enabling skills and other traits scored
Read aloud	30-35 minutes	6-7	Partial credit	Reading and speaking Oral fluency, pronunciation Content
Repeat sentence		10-12	Partial credit	Listening and speaking Oral fluency, pronunciation Content
Describe image		6-7	Partial credit	Speaking Oral fluency, pronunciation Content
Re-tell lecture		3-4	Partial credit	Listening and speaking Oral fluency, pronunciation
				Content
Answer short question		10-12	Correct/ incorrect	Listening and speaking Vocabulary
Summarize written text	20-30 minutes	2-3	Partial credit	Reading and writing Grammar, vocabulary Content, form
Write essay	20-40 minutes	1-2	Partial credit	Writing Grammar, vocabulary, spelling, written discourse Content; development, structure and coherence; form, general linguistic range

Appendix 4. TOEFL Speaking Test Rubrics. https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf

ETS TOEFL

iBT/Next Generation TOEFL Test
Independent Speaking Rubrics (Scoring Standards)

Score	General Description	Delivery	Language Use	Topic Development
4	**The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse. A response at this level is characterized by all of the following:**	Generally well-paced flow (fluid expression). Speech is clear. It may include minor lapses, or minor difficulties with pronunciation or intonation patterns, which do not affect overall intelligibility.	The response demonstrates effective use of grammar and vocabulary. It exhibits a fairly high degree of automaticity with good control of basic and complex structures (as appropriate). Some minor (or systematic) errors are noticeable but do not obscure meaning.	Response is sustained and sufficient to the task. It is generally well developed and coherent; relationships between ideas are clear (or clear progression of ideas).
3	**The response addresses the task appropriately, but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression though it exhibits some noticeable lapses in the expression of ideas. A response at this level is characterized by at least two of the following:**	Speech is generally clear, with some fluidity of expression, though minor difficulties with pronunciation, intonation, or pacing are noticeable and may require listener effort at times (though overall intelligibility is not significantly affected).	The response demonstrates fairly automatic and effective use of grammar and vocabulary, and fairly coherent expression of relevant ideas. Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. This may affect overall fluency, but it does not seriously interfere with the communication of the message.	Response is mostly coherent and sustained and conveys relevant ideas/information. Overall development is somewhat limited, usually lacks elaboration or specificity. Relationships between ideas may at times not be immediately clear.
2	**The response addresses the task, but development of the topic is limited. It contains intelligible speech, although problems with delivery and/or overall coherence occur; meaning may be obscured in places. A response at this level is characterized by at least two of the following:**	Speech is basically intelligible, though listener effort is needed because of unclear articulation, awkward intonation, or choppy rhythm/pace; meaning may be obscured in places.	The response demonstrates limited range and control of grammar and vocabulary. These limitations often prevent full expression of ideas. For the most part, only basic sentence structures are used successfully and spoken with fluidity. Structures and vocabulary may express mainly simple (short) and/or general propositions, with simple or unclear connections made among them (serial listing, conjunction, juxtaposition).	The response is connected to the task, though the number of ideas presented or the development of ideas is limited. Mostly basic ideas are expressed with limited elaboration (details and support). At times relevant substance may be vaguely expressed or repetitious. Connections of ideas may be unclear.
1	**The response is very limited in content and/or coherence or is only minimally connected to the task, or speech is largely unintelligible. A response at this level is characterized by at least two of the following:**	Consistent pronunciation, stress, and intonation difficulties cause considerable listener effort; delivery is choppy, fragmented, or telegraphic; frequent pauses and hesitations.	Range and control of grammar and vocabulary severely limit (or prevent) expression of ideas and connections among ideas. Some low-level responses may rely heavily on practiced or formulaic expressions.	Limited relevant content is expressed. The response generally lacks substance beyond expression of very basic ideas. Speaker may be unable to sustain speech to complete the task and may rely heavily on repetition of the prompt.
0	Speaker makes no attempt to respond OR response is unrelated to the topic.			

Copyright © 2004 by Educational Testing Service. All rights reserved.

Appendix 5. IELTS Speaking Assessment Criteria. Source: www.ielts.org

IELTS SPEAKING: Band Descriptors (public version)

Band	Fluency and coherence	Lexical resource	Grammatical range and accuracy	Pronunciation
9	• speaks fluently with only rare repetition or self-correction; • any hesitation is content-related rather than to find words or grammar • speaks coherently with fully appropriate cohesive features • develops topics fully and appropriately	• uses vocabulary with full flexibility and precision in all topics • uses idiomatic language naturally and accurately	• uses a full range of structures naturally and appropriately • produces consistently accurate structures apart from 'slips' characteristic of native speaker speech	• uses a full range of pronunciation features with precision and subtlety • sustains flexible use of features throughout • is effortless to understand
8	• speaks fluently with only occasional repetition or self-correction; hesitation is usually content-related and only rarely to search for language • develops topics coherently and appropriately	• uses a wide vocabulary resource readily and flexibly to convey precise meaning • uses less common and idiomatic vocabulary skilfully, with occasional inaccuracies • uses paraphrase effectively as required	• uses a wide range of structures flexibly • produces a majority of error-free sentences with only very occasional inappropriacies or basic/non-systematic errors	• uses a wide range of pronunciation features • sustains flexible use of features, with only occasional lapses • is easy to understand throughout; L1 accent has minimal effect on intelligibility
7	• speaks at length without noticeable effort or loss of coherence • may demonstrate language-related hesitation at times, or some repetition and/or self-correction • uses a range of connectives and discourse markers with some flexibility	• uses vocabulary resource flexibly to discuss a variety of topics • uses some less common and idiomatic vocabulary and shows some awareness of style and collocation, with some inappropriate choices • uses paraphrase effectively	• uses a range of complex structures with some flexibility • frequently produces error-free sentences, though some grammatical mistakes persist	• shows all the positive features of Band 6 and some, but not all, of the positive features of Band 8
6	• is willing to speak at length, though may lose coherence at times due to occasional repetition, self-correction or hesitation • uses a range of connectives and discourse markers but not always appropriately	• has a wide enough vocabulary to discuss topics at length and make meaning clear in spite of inappropriacies • generally paraphrases successfully	• uses a mix of simple and complex structures, but with limited flexibility • may make frequent mistakes with complex structures though these rarely cause comprehension problems	• uses a range of pronunciation features with mixed control • shows some effective use of features but this is not sustained • can generally be understood throughout, though mispronunciation of individual words or sounds reduces clarity at times
5	• usually maintains flow of speech but uses repetition, self correction and/or slow speech to keep going • may over-use certain connectives and discourse markers • produces simple speech fluently, but more complex communication causes fluency problems	• manages to talk about familiar and unfamiliar topics but uses vocabulary with limited flexibility • attempts to use paraphrase but with mixed success	• produces basic sentence forms with reasonable accuracy • uses a limited range of more complex structures, but these usually contain errors and may cause some comprehension problems	• shows all the positive features of Band 4 and some, but not all, of the positive features of Band 6
4	• cannot respond without noticeable pauses and may speak slowly, with frequent repetition and self-correction • links basic sentences but with repetitious use of simple connectives and some breakdowns in coherence	• is able to talk about familiar topics but can only convey basic meaning on unfamiliar topics and makes frequent errors in word choice • rarely attempts paraphrase	• produces basic sentence forms and some correct simple sentences but subordinate structures are rare • errors are frequent and may lead to misunderstanding	• uses a limited range of pronunciation features • attempts to control features but lapses are frequent • mispronunciations are frequent and cause some difficulty for the listener
3	• speaks with long pauses • has limited ability to link simple sentences • gives only simple responses and is frequently unable to convey basic message	• uses simple vocabulary to convey personal information • has insufficient vocabulary for less familiar topics	• attempts basic sentence forms but with limited success, or relies on apparently memorised utterances • makes numerous errors except in memorised expressions	• shows some of the features of Band 2 and some, but not all, of the positive features of Band 4
2	• pauses lengthily before most words • little communication possible	• only produces isolated words or memorised utterances	• cannot produce basic sentence forms	• Speech is often unintelligble
1	• no communication possible • no rateable language			
0	• does not attend			

IELTS is jointly owned by the British Council, IDP: IELTS Australia and Cambridge English Language Assessment. Page 1 of 1

Appendix 6. Machine-Human Scoring Correlation, Source: Streeter et al (2011)

Autoscoring Performance					
Response	**Assessment Prompt Material**	**N**	**Machine-Human** Score Correlation	**Human-Human** Score Correlation	**Source**
Written	81 published essay prompts (grade 6-12)	400	0.89	0.86	Prentice Hall
Written	18 research-leveled essay prompts (grade 4-12)	635	0.91	0.91	MetaMetrics
Written	5 synthesizing memos from multiple sources	1239	0.88	0.79	Council for Aid to Education
Spoken	2000 spoken English tests adults, diverse items types	50	0.97	0.98	Balogh & et al. (2005)
Spoken	3000 spoken Arabic (diverse item types)	134	0.97	0.99	Bernstein et al. (2009)
Spoken	9 Oral Reading Fluency passages for 1st – 5th grade	248	0.98	0.99	Downey et al. (2011)

There is an Open Access article, distributed under the term of the Creative Commons Attribution–Non Commercial 4.0 International (CC BY-NC 4.0) (https://creativecommons.org/licenses/by-nc/4.0/), which permits remixing, adapting and building upon the work for non-commercial use, provided the original work is properly cited.