Speaking assessment: Evolving and adapting to a changing world
Evelina D Galaczi, Cambridge English Language Assessment

The evolution of the assessment of speaking ability throughout the last century has been shaped by two competing forces – one expansive and the other reductionist. Theoretical developments in the fields of second/foreign (L2) language pedagogy and assessment, as well as external socio-political factors, have played an expansive role vis-à-vis the conceptualisation of the speaking construct. Conversely, psychometric concerns and considerations of practicality have played a reductionist role in limiting the status of speaking assessment in test batteries or in narrowing down the construct. This presentation will trace the development of speaking assessment from the late 19th to early 21st century, focusing on key tests and the pedagogic, theoretical and empirical forces which have shaped them.

Pedagogic developments at the end of the 19th century
The emergence of the Direct Method and Reform Movement in classrooms in the 1870s and 1880s brought about an interest in the use of language. Key questions and assumptions shifted and the belief espoused by the Grammar-Translation Method that speaking should not be tested was transformed into: *'How should speaking be tested?'* The Certificate of Proficiency in English exam was introduced in such a theoretical climate in 1913 by the University of Cambridge Local Examinations Syndicate. This was a milestone in the evolution of the speaking construct, since the test included an obligatory speaking component which comprised a Conversation task alongside other more traditional tasks such as dictation and reading aloud. This test signalled an engagement with direct assessment of speaking, albeit at the time narrowly defined as pronunciation accuracy.

The growing role of face-to-face speaking tests in the 20th century
The embracing of speaking assessment in the UK during the first three decades of the 20th century contrasted with scepticism in the US about the theoretical desirability of testing speaking ability, largely due to the perceived subjectivity of the assessment and the practical difficulties of administering speaking tests. The socio-economic climate after World War Two, however, highlighted the need for English speakers and marked the growing role of English as a world language. This, in turn, played an expansive role regarding the conceptualisation of speaking assessment on both sides of the Atlantic and led to the growing acceptance and use of face-to-face speaking tests with a conversational component. Key manifestations of this change, e.g. the introduction of the Lower Certificate in English at Cambridge in 1939, the launch of the United States Army Specialised Training Programme in 1943, and the Foreign Service Interview (FSI) in 1952 developed to assess the readiness of American Foreign Service personnel to communicate in real-life assignments abroad, placed speaking proficiency at the core of L2 proficiency. They also introduced some key validity-

related developments, the main being an explicit conceptualisation of the speaking construct at different proficiency levels.

As face-to-face speaking tests became more widespread, considerations of reliability and rater effects (first voiced by Edgeworth in 1888) gained more importance. In the 1940s in the UK Jack Roach, one of the key figures behind the Cambridge Proficiency and Lower Certificate exams, addressed the fundamental psychometric issues of reliability and validity through his concern with keeping standard comparable across individual examiners. The FSI team faced similar issues in the 1950s, and the FSI assessment scale signalled an important psychometric approach to defining a criterion external to rater intuition and test candidature itself.

The communicative paradigm and growing awareness of test authenticity

The communicative movement in the 1970s gave face-to-face speaking test a strong pedagogic and theoretical impetus. Influential ideas regarding the role of performance vis-à-vis competence and the role of context were offered in the 1970s by Hymes (1974), Halliday (1975) and Van Dijk (1977), and paved the way for models of Communicative Competence and Communicative Language Ability (Canale and Swain 1980, Bachman 1990) in the 1980s and 1990s. These frameworks provided a theoretical basis for the speaking construct and led to an interest in assessing the functional and communicative aspect of language. Test tasks took on a real-life purpose. Theoretically and empirically, test authenticity took centre stage, aided by the introduction of qualitative research methodologies in speaking test research in the 1990s. Following the now classic appeal by Van Lier (1989) to look "inside" the language proficiency interview, i.e. to analyse the discourse produced in speaking tests, a key question emerged: *Is the interaction elicited in speaking tests sufficient to assess communicative competence?*

Paired and group tests: expanding the speaking test construct

The introduction of paired and group speaking tests in the 1990s signalled a further expansion in construct conceptualisation. The paired/group format presented a range of interaction possibilities, including peer-peer interaction tasks, where the conversational rights and responsibilities of the participants were more balanced and a wider spectrum of functional competence could be sampled. Co-construction of interaction and the dynamic two-way influence of the interlocutors, therefore, vastly broadened the construct underlying paired and group speaking tests. The speaking test construct now drew not just on communicative theoretical frameworks, but also on frameworks of interactional competence (Kramsch 1986) and pushed construct definition of speaking tests into a new and broader conceptual terrain. The conceptualisation moved beyond a view of language competence as residing within an individual to a more social view where communicative language ability and the resulting performance reside within a social and jointly-constructed context. At the same time, empirical findings indicated that individual interviewing techniques and the background variables which interlocutors (both test takers and examiners) bring to the speaking test could affect a candidate's performance. A key construct-related question emerged: *Is the variability inherent in interaction construct-irrelevant variance (and therefore to be avoided) or is it part of the speaking construct?*

The influence of technology

The 1970s saw the growing use of computer-delivered tests, which a few decades later were followed by computer-scored tests where the assessment of speaking process was completely automated. The advent of computer-based language testing highlighted a key question: *How does the delivery medium change the nature of the construct being evaluated?* On the one

hand, computer-based speaking tests were seen as addressing a variety of caveats associated with the human factor in direct assessment of speaking, such as providing uniformity of administration, as well as the practical benefit of making the costs manageable. On the other hand, they also narrowed down the conceptualisation of the speaking test construct from a socio-cognitive construct definition in face-to-face tests, where speaking is viewed both as a cognitive trait and a social interactional one (Taylor 2011), to a psycholinguistic definition which places emphasis on the cognitive dimension of speaking (Van Moere 2012).

A glimpse into the future
Throughout the last century language testers have had at their disposal a range of speaking test formats which have been useful in eliciting and assessing speaking skills. All of these formats bring their strengths and caveats and have varied applicability for different contexts. No test format is inherently superior, and the issue of *fitness for purpose* has emerged as fundamental in driving discussions and decisions about underlying test constructs. The key question informing current debates about speaking assessment, therefore, is not whether to use a certain format or not, but: *In what contexts is a speaking test fit for purpose?* This question will continue to inform future debates as new developments in technology and natural language processing research expand the existing array of speaking tests and lead to a symbiotic relationship between technology and humans in speaking assessment.

References

Bachman, L (1990) *Fundamental considerations in language testing,* Oxford: Oxford University Press.
Canale, M and Swain, M (1980) Theoretical bases of communicative approaches to second language teaching and testing, *Applied Linguistics* 1**,** 1-47.
Edgeworth, F Y (1888) The statistics of examinations, *Journal of the Royal Statistical Society* 51**,** 599-635.
Halliday, M A K (1975) *Learning how to mean,* London: Edward Arnold.
Hymes, D (1974) *Foundations in sociolinguistics: An ethnographic approach*.
Kramsch, C (1986) From language proficiency to interactional competence, *The Modern Language Journal* 70 (4)**,** 366-372.
Taylor, L (2011) Introduction, in Taylor, L (Ed.) *Examining speaking*, Cambridge: University of Cambridge ESOL Examinations Examinations/Cambridge University Press.
Van Dijk, T A (1977) *Text and context: Explorations in the semantics and pragmatics of discourse,* London: Longman.
Van Lier, L (1989) Reeling, writhing, drawling, stretching and fainting in coils: Oral proficiency interviews as conversations, *TESOL Quarterly* 23**,** 480-508.
Van Moere, A (2012) A psycholinguistic approach to oral language assessment, *Language Testing* 29 (3)**,** 325-344.