EALTA Special Interest Group Meeting

Assessing Speaking

Thursday, 24 May, 2018, 14:00 – 17:30

PROGRAMME

14.00-14.15	**Welcome**	
14.15-14.45	**John de Jong** (Language Testing Services)	History and future of technology in speaking assessment
14.45-15.15	**Vivien Berry** (British Council) Fumiyo Nakatsuhara (University of Bedfordshire), Chihiro Inoue (University of Bedfordshire), Evelina Galaczi (Cambridge Assessment English)	Video-conferencing Speaking Tests: Insights from Examiner Feedback
15.15-15.45	COFFEE BREAK	
15.45-16.15	**Daniela Marks** (TestDaF)	Assessment of Speaking Tasks in an Academic Context – Rating Scale Development
16.15-16.45	**Tziona Levi** (Ministry of Education, Israel)	A technology-based speaking test: The general learning potential for EFL
16.45-17.30	**Panel discussion** **Nivja de Jong** (Leiden University) **Veronika Timpe-Laughlin** (Educational Testing Service) **Jing Xu** (Cambridge Assessment English)	

Opening Talk: History and future of technology in speaking assessment

John de Jong (Language Testing Services)

The prospect of using technology in the assessment of spoken language dates back to the 40s of last century, shortly after magnetic tape recorders became available. The idea then was to support rater training by using samples recorded on phonograph. But it was never implemented. Like all further developments in harnessing technology to support the assessment of speaking, it was aimed at reducing the subjective elements. Language testing in general is a relatively late discipline, and early language testing shied away from testing speaking because of distrust in the possibility to conduct it reliably, to exclude subjectivity of judgement, or to realize standardization.

It took until 1980 when finally, the administration of speaking tests was standardized by making use of pre-recorded questions using a special kind of recorder which could record the test takers' responses on a separate track. But the evaluation and scoring were still entirely manual.

Then within a few years, in 1984, the first automated scoring of speech was successful. But it took another 10 years before speech recognition techniques enabled rating speaking exams beyond the evaluation of pronunciation and speech rate and then 15 more years before it was used in large scale, high stakes testing.

Next developments will exploit the ever-increasing power of computers to implement artificial intelligence which will allow for emulating human-to-human interaction and reach the goal of objectivity and reliability in life-like settings.

Paper 1: Video-conferencing Speaking Tests: Insights from Examiner Feedback

Vivien Berry (British Council),Fumiyo Nakatsuhara (University of Bedfordshire), Chihiro Inoue (University of Bedfordshire), Evelina Galaczi (Cambridge Assessment English)

In September 2017, the Council of Europe published a companion volume to the CEFR which includes an entirely new section relating to online interaction. This acknowledges that online communication is unlikely ever to be the same as face-to-face interaction and points to several differences which may impact on the speaking construct including the availability of resources shared in real time and the possibility of misunderstandings which are not corrected immediately, which is often easier with face-to-face communication.

To explore how new technologies can be harnessed to successfully deliver a video-conferencing version of an international speaking test, a 3-phase mixed-methods study, involving a total of 220 test-takers and 22 examiners, was carried out to investigate the equivalence of the speaking construct in face-to-face and video-conferencing delivery modes of the test. The comparability of test scores and elicited language functions on the face-to-face version and video-conferencing version has generally been established (Authors, 2016, 2017a, 2017b, forthcoming). However, thematic analyses of examiner feedback collected through questionnaires and focus groups in Phases 2 and 3 of the study indicate that there are several issues to be addressed if the video-conferencing version is to be introduced.

This paper focuses on presenting such issues, in terms of examiner training, potential modifications to the examiner script and IT support. It concludes with a discussion of the affordances which video-conferencing offers for speaking assessment, and the challenges it presents which can inform decisions made by test developers who are looking to use video-conference technology for speaking tests.

Reference

Council of Europe (2017). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment; Companion Volume with New Descriptors*. Strasbourg: Council of Europe.

Paper 2: Assessment of Speaking Tasks in an Academic Context – Rating Scale Development

Daniela Marks (TestDaF)

This presentation will discuss the development of the rating scale for a speaking component of a web-based test of German as a foreign language. The target participants are international students applying to higher education institutions in Germany. The speaking section of this high stakes test consists of seven tasks and includes both isolated and integrated tasks. The responses are recorded and sent to experienced raters to be assessed with a newly developed holistic scale.

The first draft of the rating scale for the speaking component consisted of 5 criteria and 4 levels. 10 experienced raters and language testing experts used the scale for the assessment of 150 participants in a try-out of the new test. The descriptors were then revised for the first field test with 250 participants and 15 raters. This scale was then transformed into a holistic scale focussing on a variety of salient language features that seem to have the largest impact on the overall score. In addition, the responses of 250 students in the first field test of the new tasks were analysed in order to identify typical performances for each level of the scale.

This work in progress presents the steps taken so far in the development of a rating scale for speaking tasks in an academic context and gives a brief overview of some considerations regarding relevant scales.

Paper 3: Assessing academic speaking skills – rating scales as the de facto test construct

Tziona Levi (Ministry of Education, Israel)

In the past, little attention was paid to the question about the generality of students' learning potential (LP) as established with the help of Dynamic Assessment (DA) tests. The current research was conducted in the context of English as a Foreign Language (EFL) high-school exams. The research questions were: 1) To what extent a new computer-based version of oral proficiency exam is suitable for identifying students' EFL LP, and 2) Is the students' LP established with the help of DA of their oral proficiency a better predictor of their subsequent EFL reading and writing than their static oral score? 80 students (38 boys, 42 girls) received DA of their EFL oral proficiency in a pre-test – mediation – post-test format. Six months later the same students took a standard EFL reading and writing exam. The results indicate that the computer-based oral exam is suitable for the DA purposes. The mediation produced significant gain (d = 0.86) and generated a sufficiently wide range of LP scores. The correlations between oral LP scores and both reading (r=0.42) and writing (r=0.45) are significant and much stronger than the correlations with the static oral pre-test. Oral LP scores explain about 20% of the variance of reading and writing scores. There are promising initial results for research of this kind. Due to scope and time constraints, the presentation will focus on the first research question, the research design and results as they pertain to the underlying construct of the large-scale, high-stakes speaking test in the Israeli school context.