BRITISH COUNCIL
Summary of research report
ASSESSMENT RESEARCH GROUP

TITLE

Scoring validity of the Aptis Speaking test: Investigating fluency across tasks and levels of proficiency

This is a summary of a report by Parvaneh Tavakoli, Fumiyo Nakatsuhara and Ann-Marie Hunter as part of the ARAGs Research Online Series. For a copy of the full report, see www.britishcouncil.org/exam/aptis/research/publications/

WHAT WE LOOKED AT:

Fluency has for a long time been considered a key construct in second language communicative ability, a reliable indicator of L2 proficiency, and a characteristic of speech that affects listeners' perceptions of a speaker's language ability. Fluency is commonly referred to as a key performance criteria in standardised language examinations (e.g. Aptis, Cambridge General English exams, IELTS, TOEFL iBT). Despite its importance in L2 assessment, fluency is seldom investigated in sufficient detail to be fully useful for developing and validating descriptions of how the speech of L2 test-takers at different levels of proficiency differs in terms of fluency. Our aim in this study was therefore to examine oral fluency across assessed levels of proficiency, and to see whether particular characteristics of fluency (e.g. speed of speaking; tendency to pause; occurrence of repair behaviour) distinguished different levels of proficiency and different types of task.

The study addressed two research questions:
RQ1: *How are various aspects of fluency presented across different levels of proficiency (A2, B1, B2, and C1) in the Aptis Speaking test?*
RQ2: *To what extent is test-takers' fluency affected by task design (task type, discourse type and target level)?*

HOW WE DID IT:

To achieve our aim, we transcribed and analysed 120 Aptis Speaking task performances from 32 test-takers in terms of 3 measures of *speed fluency* (rate of speaking), 12 measures of *breakdown fluency* (pausing length, type, location and frequency) and 4 measures of *repair fluency* (repeating, correcting and reformulating). The micro-analyses of these measures were performed with the specialist software, PRAAT.

WHAT WE FOUND:

For RQ1, the most important findings include:

- Speed fluency (rate of speaking) distinguishes A2, B1 and B2 levels reasonably consistently. However, B2 and C1 levels are usually not different in terms of speed fluency. In other words, speed of speech is higher at B1 level than A2, and higher again at B2 level, but C1-level test-takers do not speak faster than B2-level test-takers.

- A2 speakers produce much longer silent pauses than speakers at higher proficiency levels (B1, B2 and C1).
- The two lower level speakers (A2 and B1) of the Aptis Speaking test pause at a mid-clause location more frequently than the two higher proficiency levels (B2 and C1).
- A2 speakers' use of filled pauses (e.g. uh, mmm) is very limited compared to the rest of the three proficiency groups. B1 and C2 speakers tend to use more filled pauses than B2 or A2 speakers.
- A2 speakers rarely repair their speech, while B1 speakers actively reformulate their speech. B2 and C1 speakers engage in repairs more moderately than B1 speakers.
-

These findings are encouraging as the Aptis Speaking test can utilise the above fluency characteristics as criterial features of each band level, in order to validate or modify the fluency rating descriptors of the test. However, a concern was raised in relation to the difficulty in differentiating B2 and C1 candidates in terms of their fluency performance. While the results indicated some straightforward fluency characteristics that can differentiate A2 from B1, B1 from B2, the results failed to identify a useful measure to distinguish B2 and C1 performances. One possible way to interpret this is that what makes C1 candidates different from B2 candidates is, for example, the use of more sophisticated vocabulary and complex grammatical structures rather than how 'fluent' they are. Another interpretation is that the Aptis Speaking test which has a B2 task (Task 4) but lacks a C1 task is not capable of pushing B2 and C1 candidates to their linguistic limit for fluency. The lack of a more demanding task at C1 might therefore be preventing the test from capturing differential fluency performances that could be elicited from B2 and C1 candidates.

For RQ2, a summary of the findings suggests that:

- Speed of performance is not affected by task type.
- Length of pauses is not affected by task type.
- Frequency of pauses is not affected by task type.
- Repair measures distinguish Task 3 from Task 1. Task 3 elicits most repairs.

These results imply that the performance is, in general, not affected by task type. Given the literature on the impact of task design on elicited fluency features, this finding was rather surprising and counter-intuitive. However, this may imply that the four Aptis tasks are not distinctive enough to impose different types of demand on the candidates' cognitive processes to affect their fluency performance. However, this does not invalidate the Aptis Speaking test or its by-part rating system (i.e. each part of the test is rated separately). This simply indicates that the three different scales in the Aptis Speaking test and the rating system are useful not because the tasks elicit different types of fluency performance but because they elicit different levels of fluency performance, making it easier for examiners to focus on narrower boundaries in making judgements. The use of the common scale between Task 2 and Task 3, both of which target the B1 level, is therefore justified.