Report on the

Review of draft proposal for revised specification of OET Listening

Meeting commissioned by
Cambridge English Language Assessment

Participants:

Ute Knoch (Chair)
Jonathan Silverman
John Pill
Tim McNamara

Date:

23 June 2014

Purpose of the meeting (from Cambridge English Language Assessment):
To evaluate and report on the first draft of a proposed revised specification for the OET Listening sub-test. The proposal has been drawn up by external consultants Nick Kenny and Judith Wilson in response to a commission from Cambridge English Language Assessment to produce a draft for discussion.

Rationale for proposing a revised specification (from Cambridge English Language Assessment):
The current Listening sub-test has very good face validity, and candidates and other stakeholders regard the test as authentic and congruent with listening in the workplace. However, in recent studies (Frost 2013, Macqueen et al 2013) the LTRC has identified areas for improvement which it would be worthwhile to address, such as a lack of simulation of professional-professional interaction. In the light of ongoing work to include clinical communication skills within the Speaking sub-test, there is also a need to make sure that the Listening and Speaking sub-tests continue to complement each other.

A review of the current sub-test is also an opportunity to consider how to include task types which lend themselves more readily to the consistent production of items with desirable measurement characteristics, and which facilitate the introduction of the Rasch scale we need to ensure version equivalence over time. Large-scale pretesting will be an integral part of this, and it is therefore a good time to consider alternative test designs which allow pretest items to be embedded in live versions.

It is a priority to ensure that any changes to the Listening sub-test we consider should maintain the best possible balance of validity and reliability and should build on existing strengths.

Structure of meeting/overview of day:

The meeting was held on the campus of the University of Melbourne. The participants first did an OET listening sub-test taken from the OET website to ensure all participants were thoroughly familiar with the current tasks and response formats. Half of Part A and half of Part B were completed under exam conditions.

JS reported about a recent conversation with Lesley Hay and Gad Lim (Cambridge English Language Assessment) in which it was indicated that the draft proposal for the revised listening test is very much up for discussion and open to changes.

The participants discussed the draft proposal for the revised listening sub-test in comparison to the existing test format and then focussed on the discussion questions proposed by Cambridge English Language Assessment.

For the purpose of this report, we will use the following terminology:

- Existing Part A: current Part A note-taking task
- Existing Part B: current Part B mini-lecture task
- Proposed Parts 1 to 3: new sections proposed by consultants to Cambridge ELA

In the discussion notes below, participants will at times be referred to by their initials (Jonathan Silverman = JS, Tim McNamara = TMcN).

Summary of discussion:

<u>Comparison of existing Part A with proposed Part 2</u>

The participants first discussed the existing Part A in relation to the proposed Part 2 section.

JS noted that health professionals might take notes in various ways and would have found more specific instructions as to how to take notes (how much and what to write) helpful in the existing Part A instructions. In particular, he was concerned that although the instructions said to take notes in note form, it was not clear whether to try to report items heard or a medical précis as would be found in real notes.

The participants generally liked the idea of including two dialogues, one focussing on information-gathering and one on information-giving, but there was some concern at the degree of the shortening of the texts. TMcN commented that part of the construct of this task is to measure candidates' ability to cope with time-pressured and sustained listening over a longer period. This involved speed and selection, both of which are highly authentic skills in health professionals' workplaces. Shortening the task too much might remove this important aspect of the current construct. JS agreed with this sentiment but also thought that slight shortening would be appropriate. JS also thought that dividing a whole consultation into two separate elements mirrored current practice as full consultations are not the reality for many of today's workplaces. The participants all understood that the revised specifications were designed to reduce the marking load and would result in a gain in efficiency and potentially reliability, but there was a strong concern that the proposed response format in proposed Part 2 (i.e., highly structured note-taking) would result in a major loss in authenticity and validity and reduce the specific-purpose characteristics of the OET listening test. JS commented that the proposed response format would appear to be more prone to measuring immediate recall rather than any deeper understanding, judgement or higher-level processing than the existing format rather than less. He also felt that the test was not really contextualised: as it was testing only immediate recall, it could be taken by anyone whether

a medical professional or not. All participants were concerned about the overall loss in authenticity created by the revised response format. However, the participants found it difficult to give full feedback on the proposed specifications for Part 2 because a tape script was not available, which made it difficult to comment on what the structured note-taking response format was intended to, and would actually, test. There was also insufficient information to judge what kinds of information would be targeted in the structured note-taking (including whether the grammatical accuracy of provided information was to be assessed) and whether these types of items were relevant to the professional context of the test takers.

Discussion around proposed Part 1

Positive comments focussed on the fact that it added useful variation to the current test, and that the inclusion of a variety of interactions, in particular of health professionals with other professionals, family members and caregivers, expanded the current construct of the test. However, all participants felt that there were several shortcomings: (1) the draft dialogues felt very scripted, de-contextualised, artificial and inauthentic and (2) the sample items seemed narrow in range and only testing very specific skills. It was not entirely clear to us what Part 1 is designed to measure and while everyone understands the need for the inclusion of a different response format for large-scale pre-testing, participants felt that this came at great cost to what the test is intended to measure and substantially limited the strength of the current test format (the specific-purpose nature of the test). JS felt that the proposed Part 1 does not necessarily represent what is important in a test for health professionals and that the examples of items provided could be accomplished by test-takers in any (general-purpose) language test.

Possible suggestions for changes to the proposed Part 1 given by the panel were:

- Short authentic interactions between 'real' health professionals
- Dialogues slightly longer than proposed in draft specifications
- 2-3 items per dialogue
- No more than 3-4 scenarios in total

Discussion around proposed Part 3

The panel was generally open to the idea of shortening the existing Part B, but there was concern about the proposed response format (entirely MCQs). JS was concerned that restricted item types might tap into only a few more superficial and less complex listening skills. The group thought that the variety of response types currently used in Part B was one of the strengths of the test and that this would be another instance of loss of authenticity (and therefore the specific-purpose nature of the test) in the quest for psychometric and operational gains.

Participants explained to JS the issue of variation in listening markers' performance and the need for careful training to restrict any construct-irrelevant variance introduced.

The possible inclusion of a wider variety of accents in the listening test was also discussed. A revision of the test would be a good opportunity to do this.

The reviewing team then worked through the discussion questions provided.

Response to questions posed by Cambridge English Language Assessment

Question set 1:

To what extent does the proposed specification retain the strengths of the current test, as identified in the LTRC's recent impact and validity studies? i.e.

- the consultation note-taking task is readily associated by stakeholders with workplace listening
- the consultation recording is welcomed by stakeholders as authentic in terms of interaction and speed of delivery

How could the proposed specification be improved in these respects?

The team thought that the proposed Part 2 was less authentic than the existing Part A, as the revised response format is too constrained (and probably also too easy for the typical OET test taker). While shortening the consultation was seen as acceptable, there was concern that the proposed sections were too short, given that capacity for sustained listening is part of the current test construct.

The team suggested that the two dialogues (sections) included in the proposed Part 2 be made longer than was stated in the revised specifications and that the number of items relating to each dialogue be increased. The response format should be less constrained than has been proposed in order to test more than short-term recall only. In the existing Part A, test takers also need to select what to write. The team thought that the current response format could be slightly more structured perhaps by giving titles to the bullet-points, as shown in the example below:

History of injury:

- when:
- where:
- location of pain:

The marking guide could also be constructed so that markers give, for example, marks for "any 4 of the following 5 points".

Question set 2:

To what extent is the proposed specification likely to improve on the potential weaknesses of the current test, as identified in the LTRC's impact and validity studies? i.e.

- the Part A note-taking task doesn't replicate the workplace in terms of the degree of selection/transformation of information required
- the Part A consultations are too often based in general practice
- in Part A, the range of possible keys and the marking guide are a potential source of construct irrelevant variance
- the Part B lecture topics, although 50% of the test, are not all directly related to the medical field

Where the proposal would involve a balance of gains and losses, to what extent would the balance be acceptable? Is it likely to be regarded by stakeholders as acceptable? How could the balance be improved?

The participants were not clear what was meant by 'transformation of information' in the question above (does this refer to grammatical transformation?). JS commented that although the existing Part A relies heavily on test takers' ability to recall, the proposed response format in the new part 2 did this even more. The team again noted that it was difficult to comment further without a tape script for the sample items provided.

TMcN commented that the best way to establish the effect of the proposed change in this section would be by setting up a study and examining the different outcomes for test takers when taking two different versions of the test.

The team did not agree with the second bullet-point that existing Part A consultations are too often based on general practice. Many scenarios involve other health professions. The team also did not understand why the fourth bullet-point claimed that 50% of the existing Part B of the test is currently not related to the medical field.

Consistency issues regarding the operationalisation of the marking guide are generally accepted. While this is tightly controlled with the very experienced markers based in Melbourne, it is possible that less experienced markers in other test centres might find it challenging to prepare for and carry out this sort of marking. However, the group felt that it would be possible to tighten up the existing Part A response format by making it slightly more structured (rather than what is proposed) and that this could resolve most issues with the implementation of the marking key.

The final question above asks about the balance of gains and losses. All participants thought that the losses incurred through using the proposed test format would outweigh the gains. Possible improvements to the proposal are noted in the Recommendation section below.

Everyone thought it was difficult to comment on how stakeholders would react to changes to the test.

TMcN suggested that a revision of the test would be a good opportunity to include more background noise in the listening recordings, either by setting one handover dialogue on the phone or by including routine hospital ambient noise in one of the dialogues. Such noise is highly characteristic of health professionals' workplaces but is currently not part of the construct tested. JS also suggested that it would be a good opportunity to introduce passages where the note-taking would involve selecting only important, key information rather than writing down anything that is heard.

Question set 3:

<u>To what extent does the proposed specification address the gaps identified in the LTRC impact and validity studies?</u> i.e.

- items are required which simulate patient handovers between health professionals
- an balance of intra- and inter-professional scenarios is required which matches the workplace and provides a counterpart to the scenarios in the Speaking test
- items are required which directly target higher-level processing, especially inferring meaning, listening for decision-making, listening for the follow-up action required of the listener
- items are required which better reflect the challenges of listening in the workplace, e.g. telephone handovers, chaotic environment, information relates to unfamiliar procedures and/or patients

Would candidates be likely to see the proposed task types, task formats, testing points and scenarios as worth their while to do and to prepare for?

Where there would be a balance of gains and losses, would the balance be acceptable? Is it likely to be understood by stakeholders as acceptable? How could the balance be improved?

While the first bullet-point in fact relates to a recommendation made in a study of the speaking sub-test, everyone in the team agreed that including interaction between health professionals would be an improvement to the test. The group felt that the proposed revisions to the existing Part B, in particular the inclusion of fewer items and the use of MCQ items only, actually limited the types of listening skill that could be tested, and also exposed the test to the use of test-wiseness strategies by test takers to a much greater extent than is currently the case. The examples provided for the proposed Part 1 do not demonstrate that a range of listening skills could be tested.

Again, it was difficult to comment on potential stakeholder reactions. It is unlikely that stakeholders have a clear sense of the underlying psychometric properties of a test and they are consequently more likely to focus more on the nature of the tasks and items. Therefore, it was felt that the proposed revisions might be perceived as negative (although the inclusion of professional to professional discourse would be seen as positive).

Everybody agreed with the final bullet-point proposed by Cambridge Assessment above.

Question set 4:

Is the proposed Listening test in line with the inclusion of clinical communication skills within the Speaking assessment scales? i.e.

- To what extent does it match and reinforce the enhanced Speaking construct? How can it be improved in this respect?
- Is it likely to provide examples of good practice in clinical communication? If a candidate uses a listening script as a model for their own speaking, will this be beneficial in terms of performance on the test and in terms of being ready for the workplace?
- What guidance and support would test writers without a medical background need in order to produce material which supports the clinical communications component?

JS felt that it was difficult to answer this question as there was no script provided for the consultation in the proposed Part 2. The short dialogues given for the proposed Part 1 were not really examples of communication skills. It was important that the consultation listen to good model effective communication skills and this should be taken into account by test writers. JS felt that the existing Part A test that we had listened to provided a perfectly reasonable model of a health professional's use of clinical communication skills in interaction with a patient.

The discussion then centred around guidance for test writers. The group noted that test materials are currently developed from a simulated interaction, which is audio recorded first and then manipulated as necessary. Therefore, guidance regarding content and structure only needs to be given to the health professionals involved in the recording (as is currently the case) while the item writers work from the existing audio. Therefore, no content-related guidance needs to be given to item writers. JS felt that this method of test development should be retained, as scripted dialogues did not feel authentic. He felt that candidates can

learn from materials in the current format about clinical communication skills. However instructions would need to be given to health professionals before the recording to include certain items of communication such as summarising if they otherwise failed to materialise

Question set 5:

Is the proposed test congruent with workplace tasks in appropriate ways?

- Is the link between the task types and listening in the workplace likely to be clear and satisfactory to candidates? To their future colleagues and supervisors? To representatives of regulatory bodies?
- Are there other scenarios / topics / interaction types representative of / required by the workplace which should be included in the Listening test? If so, how might they be included and what proportion of the test should they represent?
- In the proposed Part 2 note completion task, what degree of paraphrase or transformation of information would be appropriate?
- Is preparing for a test along the lines of the proposal likely to be broadly beneficial in terms of preparing for the workplace (i.e. language skills, acculturation, confidence)? Is it likely to be seen this way by stakeholders? How could it be made more beneficial in this respect?

The participants decided to differentiate between input sources and response formats when discussing this question. They thought that while they liked the inclusion of professional to professional interaction, the proposed tasks were generally less congruent with the work place than those in the existing test. While the input (if produced authentically in the same way as currently practiced) will be popular and will widen the coverage of the OET listening sub-test, the proposed response formats were highly constrained and less authentic. The team liked the different scenarios identified for the proposed Part 1 and could not think of any further important examples (although JS pointed out that there would be a myriad of sub-scenarios). The team were not able to answer the second to last bullet-point as not enough information was available.

The existing Part A requires rapid processing, paraphrasing and transforming (which, it is acknowledged, could be set up better) using a relatively dense input text. It is not clear at this stage how these skills would be tested by the constrained note-taking format proposed for Part 2.

In relation to the last bullet-point above, there was a concern that the proposed response formats (especially given the prevalence of MCQ items) would be prone to test-wiseness techniques and also strongly advantage test takers with background knowledge. The reduction in 'real' note-taking would appear to lessen authenticity and engagement with the demands of the workplace.

Overall, the group noted that the current listening test is an integrated task (listening to writing) and that this aspect of the test could be developed in revisions to the OET as a whole. For example, it would be possible to have test takers listen to a handover meeting or part of a patient consultation, take notes, and then write up a discharge letter or another appropriate text for assessment according to appropriate criteria.

Question set 6:

What information and guidance would test writers without a medical background be likely to need in order to actualise this spec?

- Are these task types/scenarios/interaction patterns likely to generate a reasonable number and variety of tasks to support a healthy item bank and multiple test forms of equivalent difficulty?
- Would the proposed specification lend itself to including representative scenarios from all 12 OET professions? How could it be improved in this respect?
- How could items be crafted so that they allow candidates to demonstrate that effective listening has taken place? (i.e. not "taking dictation", not getting the answer from background knowledge alone)
- Where inferring meaning is tested, how could items be constructed which are fair to candidates from different backgrounds – e.g. a candidate understood what they heard but came via different cultural assumption to a different conclusion

The group felt that the first bullet-point above was difficult to answer as it depended on the expertise of the item writers.

In response to the second bullet-point, everyone thought that this was dealt with more effectively in the proposed format, which allowed more professions to be used within one test version. It was noted that even professions that are not normally involved in certain interaction types would most likely to be required at least to listen to and draw information from all of them during their professional lives.

The group felt that in response to the third bullet-point, it would be best to retain a greater variety of item types, rather than only focussing on the types currently proposed.

The final bullet-point is true for all globally administered tests and it was felt that Cambridge ELA probably had the most experience with dealing with this issue. It was noted that designing a globally relevant test without watering down its specific purpose nature was difficult. Different test versions could be designed for use in different countries/contexts.

Recommendations:

The team made the following recommendations about a possible format the listening test could take:

All recordings should be authentic, i.e. recorded by subject matter experts first, with items subsequently being developed by item writers.

Part 1:

No more than 3-4 dialogues with a range of response formats and each having more than one item (2-3 items). The dialogues should be slightly longer than currently proposed by Cambridge ELA to provide greater contextualisation.

Part 2:

Two dialogues, shorter than existing Part A but longer than those currently proposed, one focussing on information-giving and one on information-gathering. One dialogue could possibly simulate an interaction in a context where background noise was noticeable or it could require test takers to only note down key information from the consultation.

Response format: we recommend a more flexible response format than that proposed in the specification but one that provides more guidance to the test taker than is the case in the existing Part A (see example above).

Part 3:

Shortened version of existing Part B involving a variety of item types.