Computer delivered listening tests: a sad necessity or an opportunity?

John Field, CRELLA Institute, University of Bedfordshire

Background: Why a CBT of listening?

Technological improvements in

- Digitised sound quality
- Transmission of sound via the internet (cf phone transmission)
- Standardisation of headphone quality
- Availability and quality of listening centre facilities

Benefits

- Flexibility
- Numbers
- Minimal supervision
- Costs

- Reduction in listening anxiety?
- Greater test taker control over content?
- Individualisation of the tasks to suit test taker's level?
- Greater construct validity??

Some little discussed questions

- Can CBT pave the way for greater use of visual material?
- What are the constraints and opportunities for multi-level tests?
- Are some formats more appropriate than others?
- How best can a test deliver items on-screen?
- What type and length of recording is best suited to computer delivery?
- Is single or double play most appropriate?
- Are there problems with computerised scoring?
- Can computer delivered materials lend themselves to diagnostic testing and remedial practice?

Greater access to visual material ?

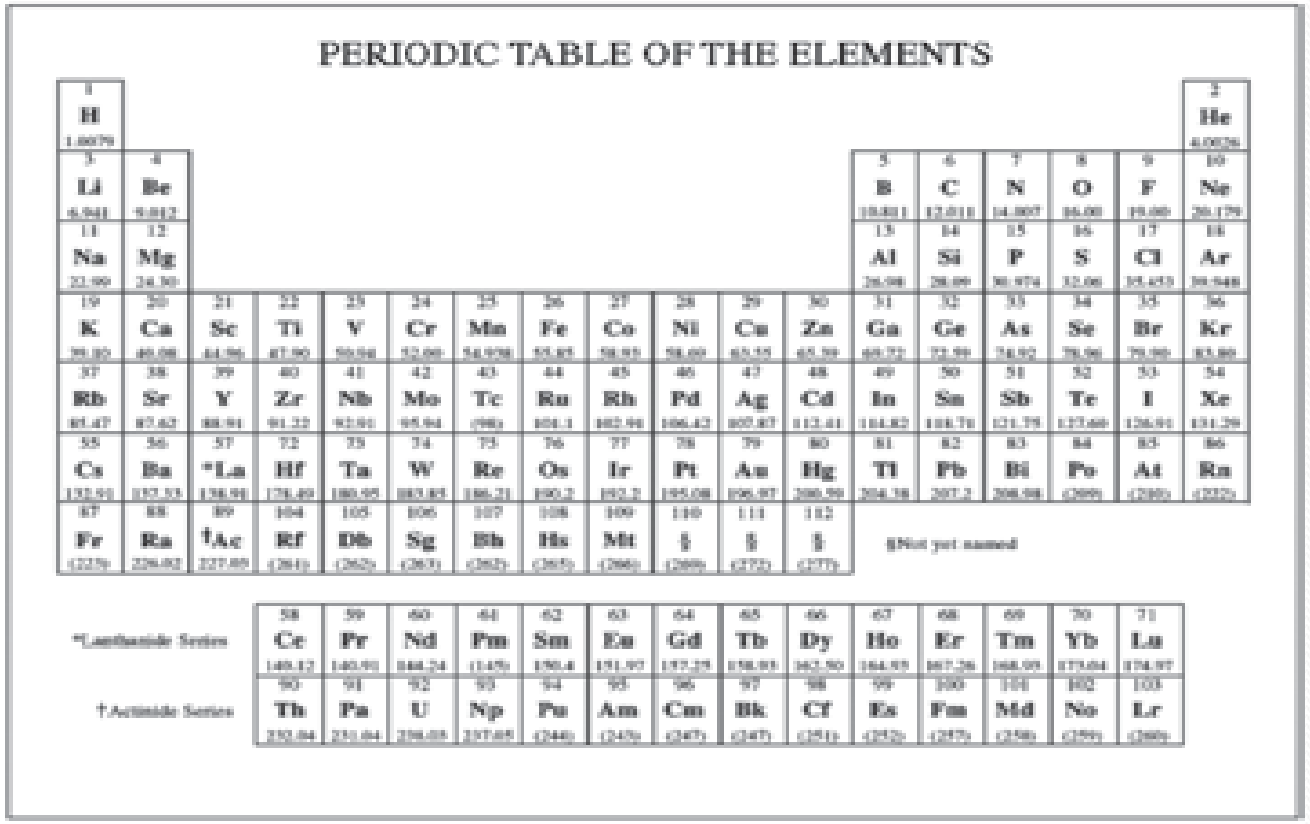
Benefits of visual support

In real-world conditions…

- A listener is assisted by paralinguistic cues from a speaker
  - Facial expressions
  - Gestures
  - Mouth movements (known to be integrated into the listening process: see e.g. McGurk,)
- There is a physical context for a dialogue
- There are probably powerpoint slides in a presentation

Visuals in CBT listening

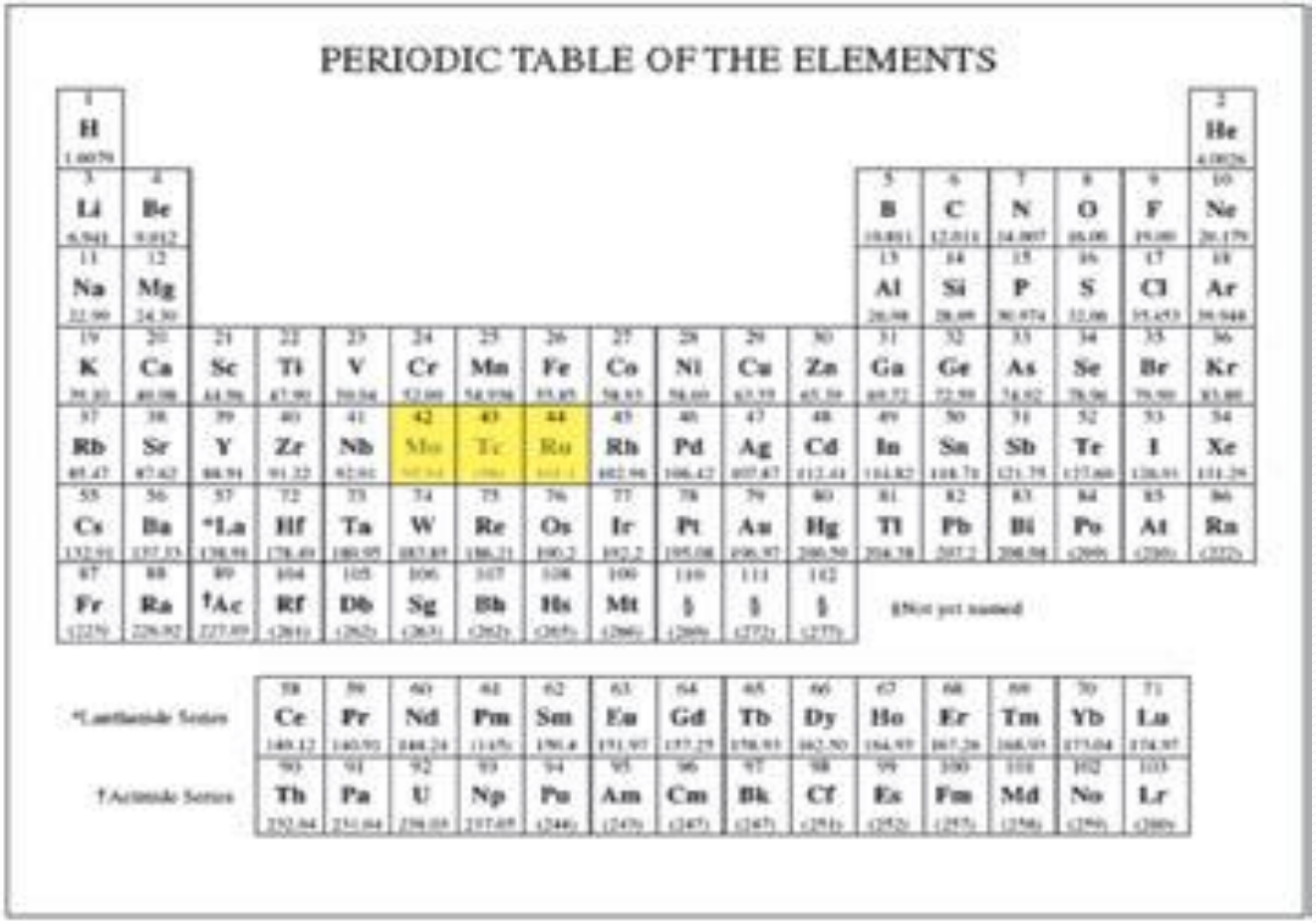
A helpful / motivational graphic?



Accompanying text

- *Um, actually, that reminds me of a good example of all this … element 43. See on the table, the symbols for elements 42 and 44? Well, in early versions of the table, there was no symbol for an element with 43 protons because an element had yet been discovered with 43 protons. So the periodic table had a gap between elements 42 and 44. And, then, uh, in 1925 a team of chemists led by a scientist named Ida Tacke claimed that they had found element 43. They had been, uh, using a relatively new technology called x-ray spectroscopy—and they were using this to examine an ore sample—and they claimed they'd found an element with 43 protons. And they named it masurium.*

A helpful / motivational graphic?



Still images

- Ginther (2002) cites evidence that test takers benefit from 'content' images which complement the information provided by the recording but not from others that just provide background context.
- Drawing on various findings, Ockey concludes (2007: 533) that listeners are not usually helped or motivated by still visual stimuli.

But for video the picture is not clear…

- The research evidence that visual information enhances L2 listening comprehension under test conditions is not as clear as might be assumed.
- Some studies (Ginther, 2002, Wagner, 2010) report clear benefits for video material over audio.
- Others (Coniam, 2001) suggest that video makes no difference
- Others (Suvarov, 2009) suggest that performance declines .
- An interpretation of the last finding is that visual information actually serves as a distraction, and that audio material focuses a listener's attention better. Or it may be that individuals differ in the use they make of the evidence provided by video, with some benefiting and others not.

Tests of academic listening

- It is anomalous that CB tests of academic listening do not aim to achieve greater cognitive validity by reproducing the real-world circumstances of listening to a lecture
- They have the opportunity of accompanying the words of the lecturer with on-screen Powerpoint slides – not simply visuals but also the text that would normally be available as a digest of the main points being made.
- There has been little research into the nature of the cognitive interaction that takes place in the mind of the academic listener between PP slides and the voice of the speaker. But see some early evidence in an eye-tracking study by Suvarov (2015).

Multi-level testing

Computer delivery of multi-level tests

- Should material be in order of difficulty or presented randomly?
- In increasing order of difficulty
- How many exponents of each level do we need to ensure reliable grading of test takers?
- I'd suggest around 8
- How do we ensure that a test taker spends most time on items at an appropriate level?
- Limit time spent on items at lower levels to ensure that higher-level candidates don't get demotivated or stop treating the test seriously.

Computational alternatives

- A. A cut off point where an individual's performance ceases to be consistent (within a margin of error)
- B. A seeding solution. According to the score on A1 items, the individual is moved on either to A2 or to B2.

  At a subsequent stage, the individual is redirected back to B1 or on to C levels.
- C. An adaptive algorithm enabling the test taker to skip sections that are too easy and finally to focus on a section at the most appropriate level.

Constraints on formats

Formats that work…

- Multiple-choice
- Visual multiple-choice
- Multiple matching
- True/false/not mentioned
- Identify three main points (not in same order as recording)
- NOT gap fill

- Tick box
- Drag and drop
- Drag and drop
- Tick box
- Tick box
- ? Drag and drop at lower levels?

Scoring: test designers beware

Test providers are quite often reliant on standard platforms. The platform may not be capable of

- Scoring in terms of location on the screen

  ?? Drag and drop??

- Scoring according to whether pieces of text have been placed in order

  ?? Identify speaker's points and place in order??

- Scoring according to whether ticks appear in correct place in a table

  ?? Multiple matching??

Delivering items on-line

- So a tester has found a riveting 3 minute recording about the life of the housefly.
- She has found 6 major points to target with 4-option multiple-choice items
- Ergo: 30 lines of text. The test taker therefore has to scroll down while listening
- Problem: This adds to the divided attention demands of needing to read while listening. It also adds to item-order pressures: 'Have I missed Q1 and should I now be listening out for Q2?' (Field, 2012)

Solutions

- The CBT mode lends itself to extensive use of short clips of 15-40 seconds followed by a single question or by two questions.
- These can be used, even at higher proficiency levels, so long as they target higher-level processes (identify main point or speaker's attitude, infer information not made explicit, etc).
- Though some longer (3 min.) recordings are obviously necessary at higher levels, MCQs should be avoided. Response formats can be simplified by the use of multiple-matching, drag & drop etc.

Single or double play?

Options

- Single play. Fails to compensate for the lack of the visual evidence or the unnatural cognitive demands of most test formats.
- Optional double play. Test taker can opt to play the recording twice.
- Optional replay. Test taker can revert to any part of the recording that has not been understood
- Double play. Recording heard twice.

The last three can be claimed to reduce the extent of listening anxiety experienced by the test taker.

Issues raised by optional replay

- The test designer has to determine maximum length for the test, allowing adequate time for replays.
- Should there be additional marks for a test taker who completes the test within a shorter time limit (i.e. with minimal replay)?
- Test takers, especially weaker ones, need to make decisions about a) how accurately the recording has been understood; b) whether sufficient time remains for a second hearing. This adds an important construct-irrelevant element to the cognitive demands of the test.
- Contrary to what has just been suggested, the stress associated with decision-making might particularly affect those suffering from listening anxiety.

A double play solution that fits CBT well

- Research by Sherman (1997) demonstrated that performance in listening tests improves when test items are delivered between the two plays of a recording.
- It is much easier to adopt this procedure in CB tests than it is in paper based ones ,where answer sheets have to be handed out.
- Advantages: It ensures that, at least during the first play, test takers listen to a recording in a way that resembles real world listening : i.e. without

  a) the prior knowledge gained from scanning *written* information in the form of pre-set test items

  b) the test-wise strategies that the availability of this information encourages.

  See Field (2015) for a full discussion.

Double play to reduce scrolling down

- Test takers hear a standard 3-4 min. recording all the way through.
- On a second play, the recording is divided into three parts. Before each part, two or three items appear on the screen. Test takers are allowed time to answer them, before the next part begins and a new set of questions appears…

Diagnostic testing

Progress tests

- In a listening centre, test takers listen to the recording once , and attempt to answer questions.
- On a second play, they are provided with a transcript. They highlight sections which they find difficult.
- This feeds into small-scale practice exercises that focus

a. on perceptual features that cause problems of word / phrase recognition

b. on syntactic and connective features that cause problems of parsing or understanding

Transcript

- *I've had quite a lot of experience of overseas travel + so I do know that it's important to get to the airport with plenty of time + and I guess that's what I should have done this morning*

Transcript

- *I've had **quite a lot of** experience of overseas travel + so I do know that it's important to get to the airport with **plenty of** time + and I guess that's what **I should've done** this morning*

Practice materials

- *I should've done*
- *He might've done*
- *They must've done*
- *I needn't've done*

- *I should've thought*
- *I should've remembered*
- *I should've paid*
- *I should've apologised*

References

- Coniam, D. (2001). The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: a case study. *System*, vol. 29: 1–14.
- Field, J. (2012a) The cognitive validity of the lecture based question in the IELTS listening paper. In L. Taylor & C. Weir et al. (eds.) *IELTS Collected Papers 2: Research in Reading and Listening Assessment*. Cambridge: Cambridge University Press: pp. 391-453.
- Field, J. (2015) The effects of single and double play upon listening test outcomes and cognitive processing. London: British Council ARAGS Reports, 2015 http://www.britishcouncil.org/exam/aptis/research/publications
- Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing*, vol. 19: 133–167.
- Ockey, G.J. (2007) Construct implications of including still image or video in computer-based listening tests. *Language Testing*, 24: 517-537
- Sherman, J. (1997) The effect of question preview in listening comprehension tests. *Language Testing*, 14, 185-213.
- Suvarov (2015). Interacting with visuals in L2 listening tests: an eye-tracking study. London: British Council ARAGS Reports, 2015. Available at:
- https://www.britishcouncil.org/exam/aptis/research/publications/interacting

Solution 1: Control recording length

- A1-B1: clips of around 15 seconds
- B1+: recordings of 30 seconds
- B2 / C levels: recordings of around 2.5 – 3 minutes
- Advantages:
- Higher level test takers don't lose motivation through spending a lot of time on easy material
- Possible to recognise a cut-off points where test takers no longer perform consistently