Comparing multiple-choice and open-ended items in listening tests: Implications for their use in an academic context

Andrew Fleck, Trinity College London

June 2024, EALTA Belfast

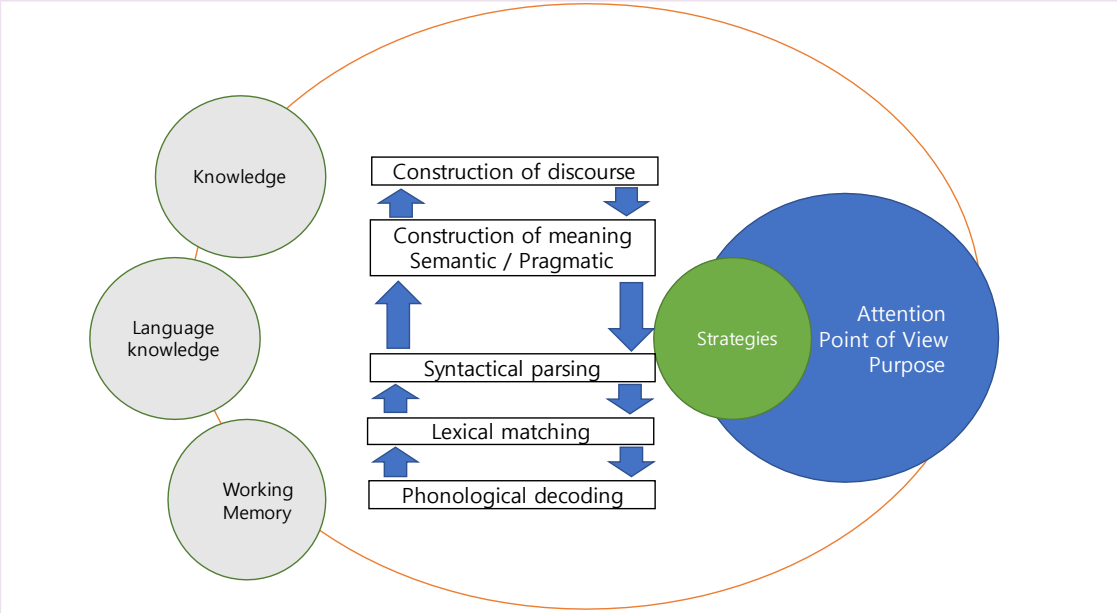
With thanks to Lancaster University and Trinity College London

Outline

- Introduction
- Item types' difficulty
- How item types affect input understanding
- How item types affect test taker's approach
- Discussion

Listening

- Weir (2005) understands processes to be at the heart of cognitive ('theory-based') validity
- Listening consists of a number of concurrent processes (see Field, 2013, Rost, 2011)
  - Lower: phonological, lexical, syntactical
  - Higher: semantic, pragmatic, discourse
  - Strategies: metacognitive, cognitive; tactics, test-wise tactics (See Vandergrift and Goh, 2021)
- Rost (2011) notes that comprehension is on a spectrum, rather than all or nothing



Note. The model combines elements from "Teaching and researching listening (2nd Ed.)," by M. Rost 2011, p. 59 Rost (2011) and "Examining listening : research and practice in assessing second language listening" by J. Field, 2013

Multiple-choice questions (MCQ)

Which actor was (briefly) a teacher in Newcastle upon Tyne?

a. Kenneth Branagh
b. Liam Neeson
c. Stephen Rea
d. Ciaran Hinds

Open ended questions (short)

Which county is the geographical centre of *Ireland* in?

_____

The geographical centre of *Ireland* can be found in the county of _____.

(not ‘Why did Liam Neeson quit teaching?’)

Comparing the items

- Are the item types comparably difficult and discriminating?
- How well must test takers understand the input before they can answer each item type?
- What range of approaches do test takers access while answering each item type? (allowing us to make inferences about levels of processing and strategies)

Methodology

- Sourced higher level MCQ testlets for a Trinity exam
- Wrote equivalent or alternative OE items for each MCQ item
  - For about half the items it was possible to write directly equivalent and quality OE items (same question, same answer)
  - Items with summarising, evaluative, inferential aspects could not be converted into OE
    - This accords with a common critique of OE items Less flexible in terms of level of processing – tests words or phrases, and local rather than global understanding(Allen et al., 1988; Field, 2012; Haladyna & Rodriguez, 2013; Shohamy & Inbar, 1991)
- OE items went through a quality review process at Trinity

Methodology: parallel tests

Set	Number	Tasks
A	33	MCQ1, MCQ2, OE3, OE4
B	27	OE1, OE2, MCQ3, MCQ4

- Four testlets, each with an MCQ and OE version, arranged as above.
- Test takers were EAP students at UK university
- Various descriptive and inferential statistics derived through analysis of test performances within and across sets

Are the item types comparably difficult and discriminating? Background

- Berne (1992) and Cheng's (2004) studies found OE more difficult than MCQ by factor of 1.6
- Innami and Koizumi's (2009) meta-analysis found MCQ have 78% chance of being easier than OE (but included longer OEs)

Are the item types comparably difficult and discriminating? Results

Within sets

- The mean scores for the 10 MCQ and 10 OE items for set A were, 5.97 and 3.61 respectively. For set B they were 5.15 and 5.11
- In the bottom third of test takers the preference for MCQ was more marked (A: 4.00, 0.72; B: 2.33; 2.00)
- In both sets, the five items with the highest facility value were MCQ and the five items with the highest discrimination were OE. Even where FV was similar discrimination was higher in OE items.

Across sets

- The equivalent items were subjected to a T-test with item type as a variable: in all but two cases the MCQ had a higher mean (many of which were significant (sig = >.05)

Are the item types comparably difficult and discriminating? Conclusions

- The MCQ items were found to be easier than the OE items, though the effect was unpredictable.
- This difference was more pronounced among lower levels
- The OE items have higher discrimination

Methodology: The qualitative instrument

- Purpose: gain access to participants thought processes during test *and* generate meaningful and processible data
- Verbal protocol was not practical
- Used *stimulated recall* in combination with a *coding questionnaire* (Nevo, 1989)
  - After the test, each recording is played section by section, each section corresponding to where an answer is found. This acts as an artefact to retrieve WM processes used during the test (Ericsson & Simon, 1993; Gass & Mackey, 2018).
  - Participants 'code' their responses to the test, providing one code related to their level of comprehension and one related to their 'approach'.

Methodology: The qualitative instrument

- 60 test takers coded 10 MCQ and 10 OE items, so there were 600 coded responses for each item type
- The Chi-square for independence indicates a significant association between correctness of response and the reported level of comprehension: X^2 (7, n = 588) = 97.541, *p* = <.001, Cramer's V = .407 for MCQ items; X^2 (7, n = 600) = 190.118, *p* = <.001, Cramer's V = .563 for OE items).
- The Chi-square for independence indicates a significant association between correctness of response and the approach used: X2 (8, n = 588) = 68.793, p = <.001, Cramer's V = .342 for MCQ items; X2 (9, n = 598) = 174.154, p = <.001, Cramer's V = .540 for OE items).

How well must test takers understand the input before they can answer each item type?
Background

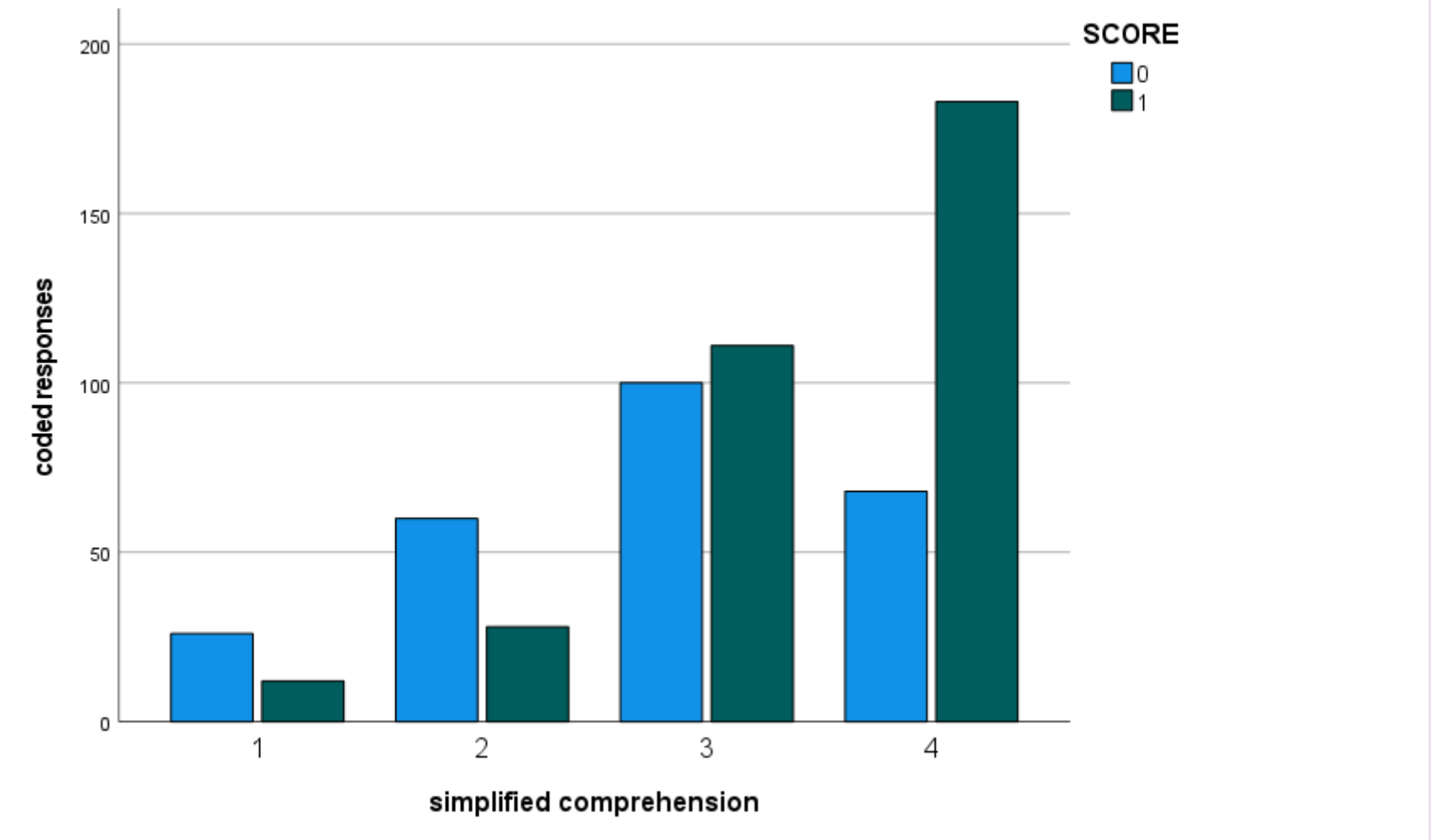
- The information in MCQ options can support listeners – providing back up to their understanding of the text (Clark and Clark, 1977 (as cited in Cheng, 2004), Kozo & Green, 2008, Vandergrift, 1998)
- MCQ represents an easier process –'confirming options' (Cheng, 2004); while OE requires test takers to rely solely on what they hear (Field, 2012)
- So we might expect test takers to understand more easily during an MCQ-based test

How well must test takers understand the input before they can answer each item type?
Background

Rost's (2011) continuum of understanding	**Continuum of understanding in this study**	**Other information**
Non-understanding	Do not understand	Did not know which part was relevant Did not catch the relevant part Did not understand the relevant part
Misunderstanding	Partially understand	
Partial understanding		
Plausible understanding	Mostly understand	But was unsure of answer
Acceptable understanding		And was sure of answer
Complete understanding	Fully understand	But was unsure of answer
		And was sure of answer

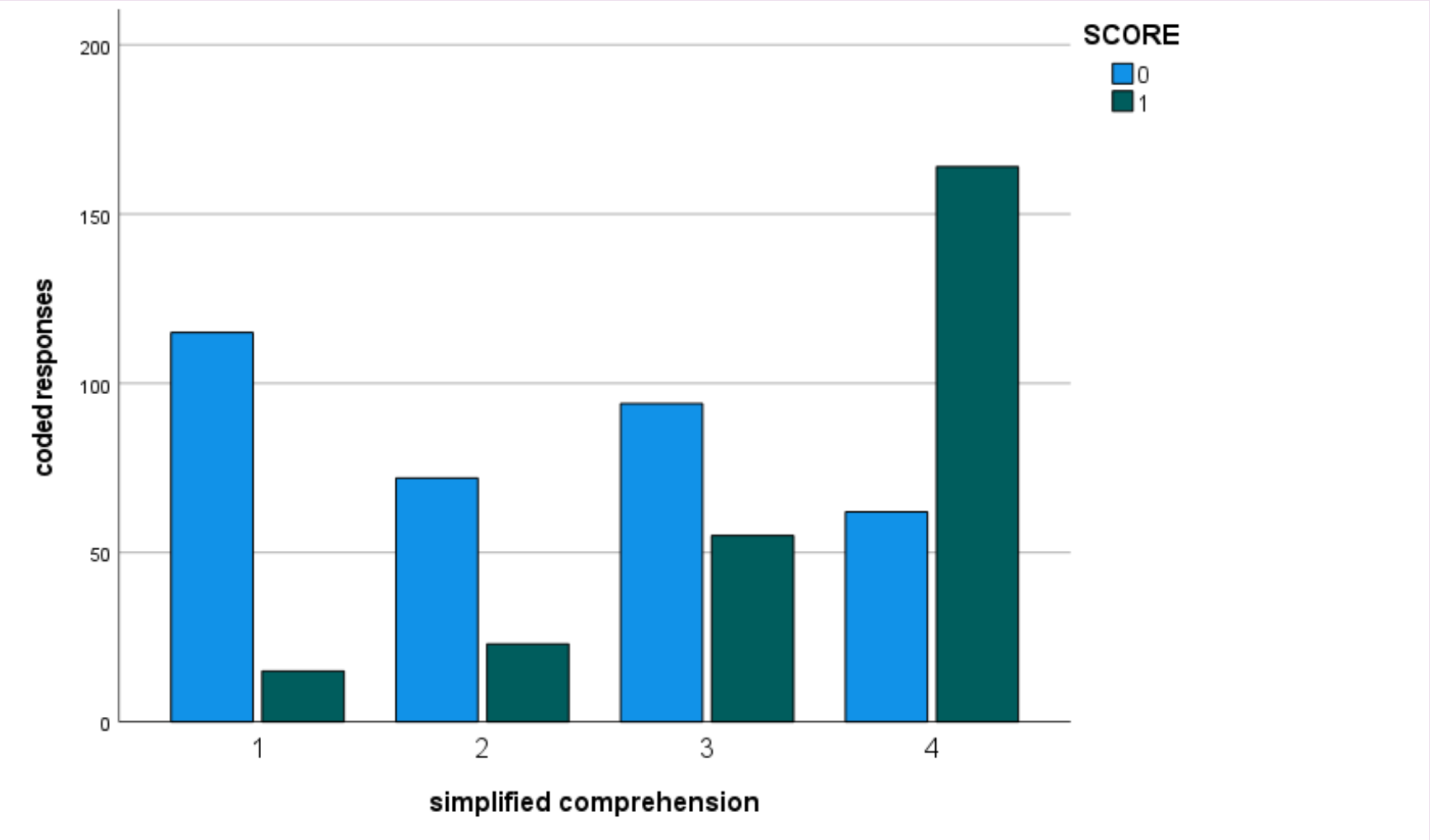
Perceived comprehension MCQ

1. Do not understand
2. Partially understand
3. Mostly understand
4. Fully understand



Perceived comprehension OE

1. Do not understand
2. Partially understand
3. Mostly understand
4. Fully understand



How well must test takers understand the input before they can answer each item type?
Results

- MCQ are easier for those who say they ‘mostly understand’ (but not those who ‘partially understand’)
- OE is more ‘all or nothing’. It is mostly those who ‘fully understand’ who get the answer. Those who mostly understand, tend not to.
- Test takers tend to report lower comprehension while doing OE tasks.

What range of approaches do test takers access while answering the items?
Background

- In MCQ items, any level of processing can be targeted (Bachman, 1990, as cited in Cheng, 2004)
- OE described as less flexible in terms of level of processing – tests words of phrases, and local rather than global understanding; narrow strategies used(Allen et al., 1988; Field, 2012; Haladyna & Rodriguez, 2013; Shohamy & Inbar, 1991)

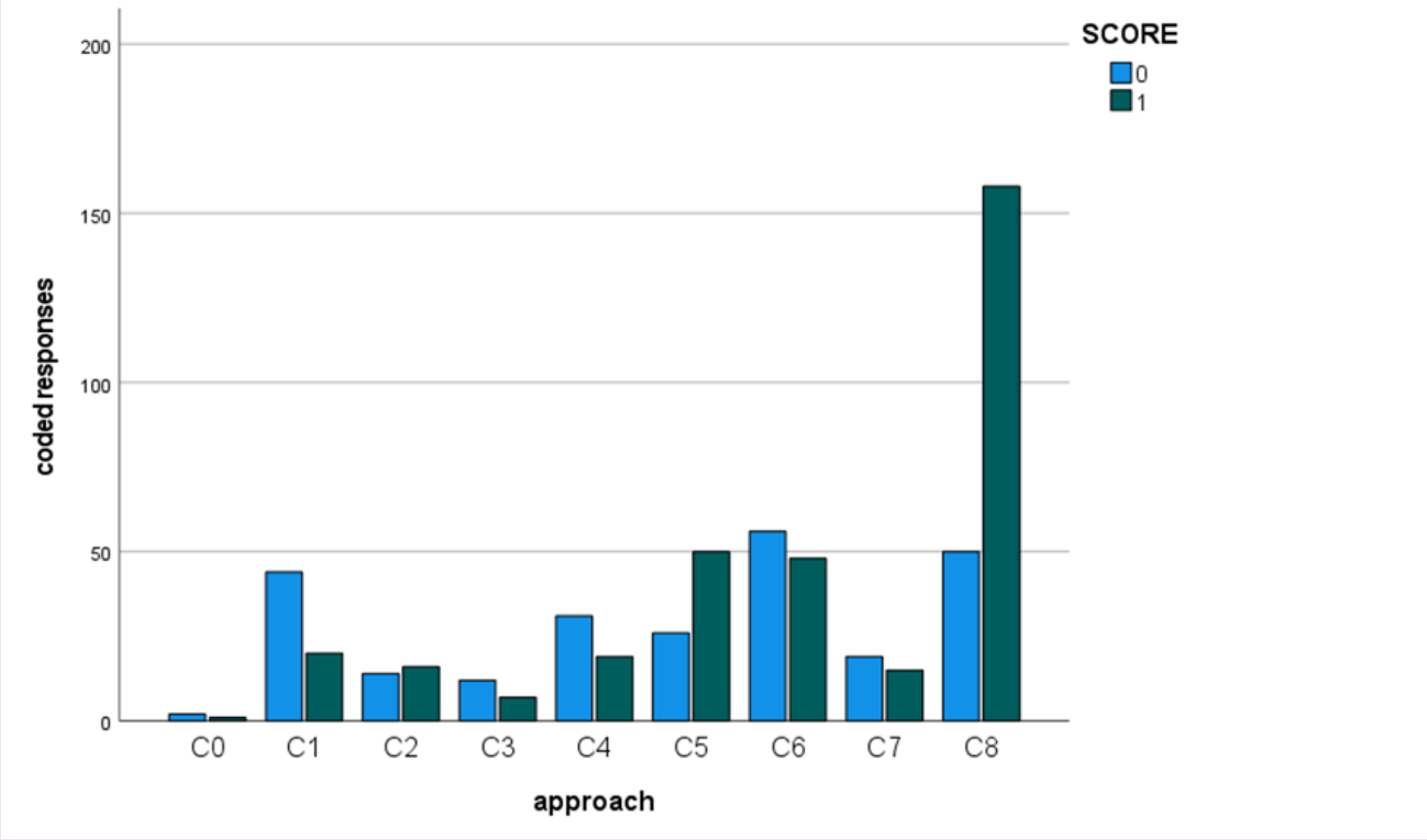
Approaches MCQ

C0/D0 I did not answer
C1/D1 I guessed
C2/D2 I answered using common sense, not based on the recording.
C3/D3 I answered using specific knowledge I already knew before the recording.
C4/D4 I answered based on words I heard that were the same as words in an option / in the question
C5/D5 I answered based on words I heard that were synonyms of words in an option / in the question
C6/D6 I answered based on some sentences that I understood.
C7/D7 Although I missed the answer, I guessed based on general understanding of what was going on or being said.
C8/D8 I answered based on my understanding of what I heard.
C9/D9 Other (please specify)

- Codes show where items engage higher or lower levels of processing, local or global understanding (Kozo & Green 2008; Nevo, 1989; Thissen, 1989; Shohamy & Inbar, 1991; Vandergrift, 1998) and particular strategies (Goh, 2002).
- At the same time I tried to use non-leading language for the test takers

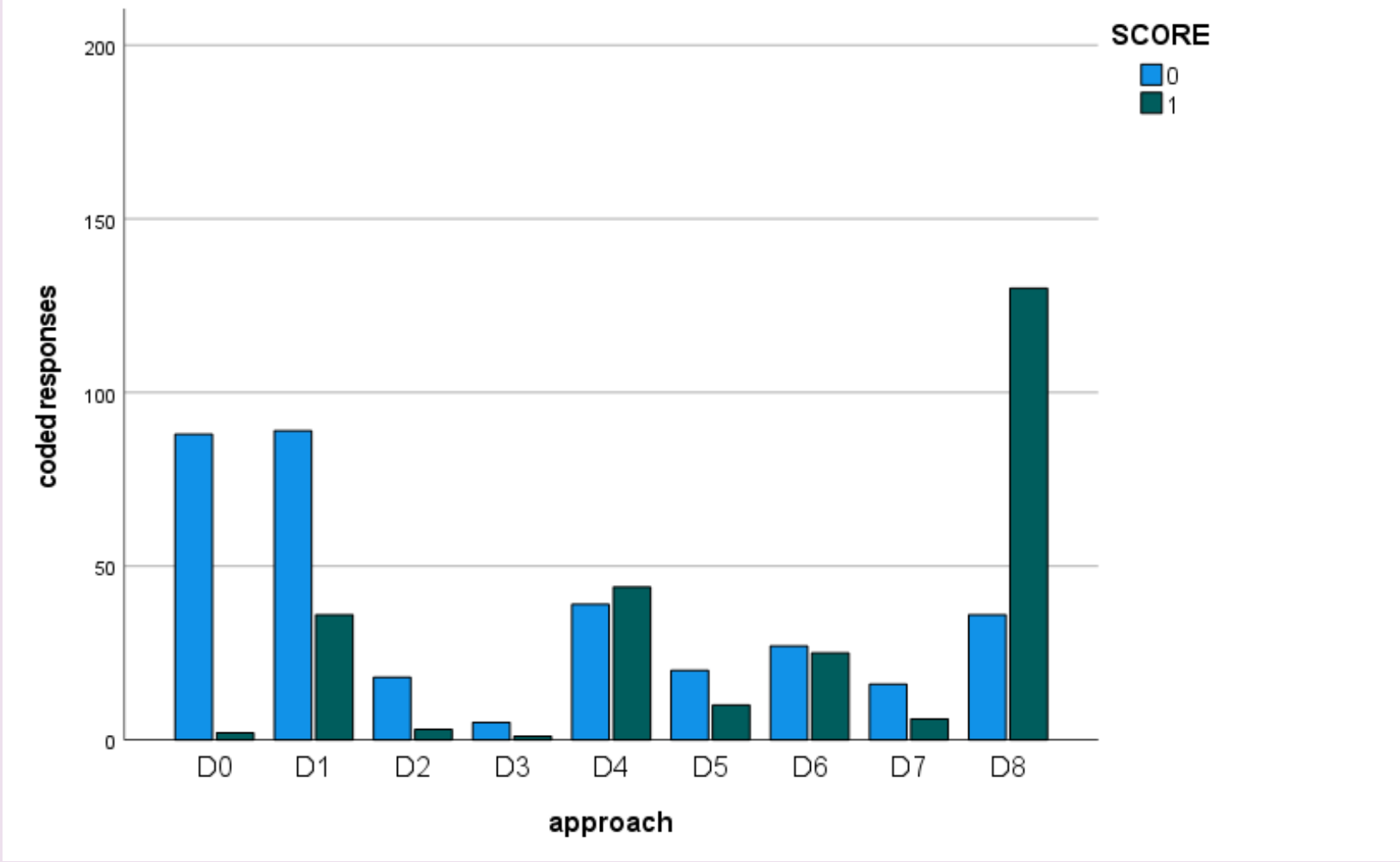
Approaches MCQ

C0/D0 I did not answer
C1/D1 I guessed
C2/D2 I answered using common sense, not based on the recording.
C3/D3 I answered using specific knowledge I already knew before the recording.
C4/D4 I answered based on words I heard that were the same as words in an option / in the question
C5/D5 I answered based on words I heard that were synonyms of words in an option / in the question
C6/D6 I answered based on some sentences that I understood.
C7/D7 Although I missed the answer, I guessed based on general understanding of what was going on or being said.
C8/D8 I answered based on my understanding of what I heard.
C9/D9 Other (please specify)



Approaches OE

C0/D0 I did not answer
C1/D1 I guessed
C2/D2 I answered using common sense, not based on the recording.
C3/D3 I answered using specific knowledge I already knew before the recording.
C4/D4 I answered based on words I heard that were the same as words in an option / in the question
C5/D5 I answered based on words I heard that were synonyms of words in an option / in the question
C6/D6 I answered based on some sentences that I understood.
C7/D7 Although I missed the answer, I guessed based on general understanding of what was going on or being said.
C8/D8 I answered based on my understanding of what I heard.
C9/D9 Other (please specify)



Results: approaches

- MCQ items elicit and support some listening strategies that OE doesn’t.
- Successful listening (whatever the task type) seems to involve fewer conscious strategies – or they are so smoothly integrated into the act of listening that the listener does not acknowledge them (c.f. Macaro et al., 2007; Peters’ 1999)

Discussion 1: decoding and meaning

- University entrance levels typically set at B2 or above. Lecturers want students to attend to what they are saying, not decode it.
- Field (2019) has emphasised attending to meaning in higher level tests. Therefore, items should be chosen that can measure ability to attend to meaning.

- Item development stage suggested that OE are limited in the types of processing that can be engaged and the subskills tested.
- MCQ items are more flexible in terms of the sorts of understanding and the processes that can be targeted.

Discussion 2: supporting understanding

- University is a multimodal learning environment. Support to understand meaning comes from a variety of sources – written materials, peers, recordings etc.

- MCQ offers support to test takers who ‘mostly’ understand. The support offered (as Cheng, 2004, argues) could be argued to be analogous to contextual knowledge in real life listening.
- OE test takers tend to feel that they have a lower comprehension of the text, perhaps because they are trying (and failing) to attend to individual words.

References

Allen, E., Bernhardt, E. B., Berry, M. T., & Demel, M. (1988). Comprehension and text genre: An analysis of secondary school foreign language readers. *The Modern Language Journal (Boulder, Colo.)*, *72*(2), 163–172. https://doi.org/10.1111/j.1540-4781.1988.tb04178.x

Berne, J.E. (1992). *The effects of text type, assessment task, and target language experience on foreign language learners' performance on listening comprehension tests*. ProQuest Dissertations Publishing. https://www.proquest.com/docview/303977181?pq-origsite=primo

Cheng, H.F. (2004), A Comparison of Multiple-Choice and Open-Ended Response Formats for the Assessment of Listening Proficiency in English. *Foreign Language Annals*, 37, 544-553. https://doi-org.ezproxy.lancs.ac.uk/10.1111/j.1944-9720.2004.tb02421.x

Ericsson, K.A., & Simon, H. A. (1993). *Protocol Analysis*. MIT Press. https://doi.org/10.7551/mitpress/5657.001.0001

Field, J. (2012). The cognitive validity of the lecture-based question in the IELTS listening paper. In L. Taylor & C. Weir (Eds.). *IELTS collected papers* 2: research in reading and listening assessment (pp. 391-453) Cambridge University Press.

Field, J. (2013). Cognitive validity. In A. Geranpayeh, & L. Taylor (Eds). *Examining listening : research and practice in assessing second language listening*. Cambridge University Press.

Field, J. (2019). *Rethinking the second language listening test : from theory to practice*. Equinox Publishing.

Gass, S.M., & Mackey, A. (2016). *Stimulated Recall Methodology in Applied Linguistics and L2 Research.* (2nd ed.). Taylor & Francis Group.

Goh, C.C.M. (2002). Exploring listening comprehension tactics and their interaction patterns. *System (Linköping)*, *30*(2), 185–206. https://doi.org/10.1016/S0346-251X(02)00004-0

Haladyna, & Rodriguez, M. C. (2013). Developing and validating test items (1st ed.). Routledge. https://doi.org/10.4324/9780203850381

References

Herbert, D., & Burt, J. S. (2003). The effects of different review opportunities on schematisation of knowledge. *Learning and Instruction*, *13*(1), 73–92. https://doi.org/10.1016/S0959-4752(01)00038-X

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*(2), 219–244. https://doi-org.ezproxy.lancs.ac.uk/10.1177/0265532208101006

Kozo, Y., & Green, A. (2008). To show or not to show: The effects of item stems and answer options on performance on a multiple-choice listening comprehension test. *System (Linköping)*, *36*(1), 107–122. https://doi.org/10.1016/j.system.2007.12.003

Macaro, E., Graham, S., & Vanderplank, R. (2007). A review of listening strategies: focus on sources of knowledge and on success. In A.D. Cohen, & E. Macaro (Eds.). *Language learner strategies : thirty years of research and practice*. Oxford University Press. https://www-vlebooks-com.ezproxy.lancs.ac.uk/Product/Index/702382?page=0&startBookmarkId=-1

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing*, *6*(2), 199–215. https://doi-org.ezproxy.lancs.ac.uk/10.1177/026553228900600206

Peters, M. (1999). *Les stratégies de compréhension auditive chez les élèves du Bain Linguistique en français langue seconde* [Doctoral thesis, University of Ottawa]. University of Ottawa Theses. https://ruor.uottawa.ca/bitstream/10393/8654/1/NQ48111.PDF

Rost, M. (2011). *Teaching and researching listening* (2nd Ed.). Harlow: Pearson. https://ebookcentral.proquest.com/lib/lancaster/detail.action?docID=5268626.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-Choice Models: The Distractors Are Also Part of the Item. *Journal of Educational Measurement*, *26*(2), 161–176. http://www.jstor.org/stable/1434863

Vandergrift, L. (1998). Successful and Less Successful Listeners in French: What Are the Strategy Differences? *The French Review*, *71*(3), 370–395. http://www.jstor.org/stable/398969

Vandergrift, L, & Goh, C. C. (2021). *Teaching and learning second language listening: Metacognition in action*. Routledge.