Testing & Evaluation & Assessment

TEASIG Webinar Series

Multiple choice items – What they are good for and how to write them
Glyn Jones 12/5/20

As time did not allow for all the participants' questions to be answered during the webinar, Glyn has provided comments and answers below.

Thank you for the excellent questions submitted during the webinar! Some of them, particularly the earlier ones, I think I answered in the course of my talk or in the Q and A afterwards – I certainly hope so and apologise if your answer is not here. Where two or more people have asked very similar questions, I have conflated them and given one answer. I have reproduced my references as there were some questions about these, and added one or two more.

Isn't it the truth that multiple-choice questions are used solely because they are very easy to correct for a large number of test takers?
No. The convenience of mechanical or clerical marking accounts for much of the attraction of MC items for test developers, but it is not the whole story. For some types of comprehension assessment, they can be more effective than open-ended questions (e.g. where these could be answered by lifting chunks from the reading passage without necessarily understanding). Also, they can be designed so as to fine-tune the difficulty level of the item and the precise construct being assessed. As for economy, the convenience of marking has to be offset against the relatively high cost of producing the items in the first place.

The item about 'lockdown' also brings to mind that test takers can know the answer from knowledge outside the test.
I don't think this applies particularly to that item (knowing what prepositions collocate with 'lockdown' *is* linguistic knowledge), but it is true that it can be difficult to write items where the test taker can't answer from knowledge of the world without having to understand the text, particularly with non-fiction (factual) texts. For example, if the text is on about climate change and one of your items has as its key "The accumulation of carbon dioxide in the atmosphere contributes to global warming", many test takers will be able to answer it without reading the text. So better to target for the less obvious points.

Is it necessary to arrange the choices starting from the shortest (A) to the longest (C)?
No, I haven't come across this recommendation. Regarding order of options, though:

- Some test-wise guides give the advice, "if in doubt, choose C", because apparently item writers tend to place the key in third position (in four-option items) more often than in the other three. This doesn't work, of course, if the order of options is randomized by a computer.
- In two recent studies involving listening comprehension items, Franz Holzknecht and colleagues found that the position of the key made a difference to the difficulty of the item: items were easier if the key was in first position than if it came further down the list of options. This argues against the practice, followed in many computer-based tests, of randomizing the order of items differently for each test taker, as it will arbitrarily favour those test takers who get to see the key in first position.

I have noticed that some teachers use exact lexical matches as the key with A1/A2 level students. What do you think of that?
At very low levels it could be that you *do* want to test the ability to locate particular words in a text. More often, though, we are interested in the ability to process meaning. As always, it depends on the construct.

What's your stance on multiple correct answers (even more than 2) – especially with regard to validity (and reliability)?
I think that multiple-answer multiple-choice items are effectively a different item type. For one thing, they are more complicated to score. You have to apply an algorithm, such as one point for each chosen key minus one point for each chosen distractor (but what if the test taker only selects distractors and no keys? Do you award a negative score?). This is no obstacle, of course, in a computer-based test, where a complicated scoring algorithm can be applied effortlessly. From the point of view of content, these items often take the form of "Which of these statements is true according to the writer?" so arguably could be replaced by a series of separate items. I don't know any of any research on the relative validity and reliability of this item type.

How would you complement MC items in scenario-based assessment?
In my understanding, scenario-based assessment typically comprises a series of related tasks that aim to resemble real-life language use, and the test taker is assessed on their achievement of the goal. It may be that, on the way, it is expedient to check whether the test taker has understood crucial points in the source material, and that MC items may be useful for this. Otherwise, I would expect that the final assessment consists of a judgment of productive skills, and MC items would play no part.

Can you use negation in the distractors or key?
Yes. If the targeted information is negative, then it would seem natural that the key is negative. For example:

> Stem: *Why did he get dressed in the dark?*
>
> Key: *Because he didn't want to wake his room-mate.*

Possible plausible distractor: *Because he couldn't find the light switch.*

The problem with negation is in items such as "Which of these things did he NOT do?" This becomes a logical puzzle in which the distractors are double negatives.

You said to avoid stems with "not". Would a stem like "What opinion does the author disagree with?" be acceptable even though it's basically asking "What opinion does the author not agree with?"
See my response above. If the targeted information is negative, then it may be natural to include negation in the stem, but still better to avoid it. To take your example, yes, I think "What opinion does the author disagree with?", where the negation is embedded in the lexis, is less likely to confuse than "What opinion does the author not agree with?"

Could you please expand on the idea that true/false items are not adequate for language tests?
I think they are fine for classroom exercises where the stakes are not high and there is scope to discuss the answers. For high stakes tests, there are two issues: the 50/50 probability of answering correctly (true of any binary choice question, of course) and the findings of some research that they test some cognitive trait other than comprehension. In one study, Grosse and Wright (1985) found that when they split a T/F test into two subtests, one with all correct answers being "true" and the other with all correct answers being "false", there was a very low correlation between scores on the two tests. They hypothesized that the more able test takers were prone to seeing something wrong with a supposedly "true" statement whereas less able test takers students more readily agree with any printed statement.

For true/false we can ask for justification to make it more challenging.
Yes, but then scoring becomes more complicated. You have to devise a mark scheme in which you specify carefully what you accept as justification. You lose the convenience of MC, so perhaps should consider using open-ended comprehension questions instead.

How about deducting various points for wrong answers?
I think this would complicate the mark scheme for no benefit. I have heard (but can't recall having read) of the suggestion that this could be used as a way to discourage guessing, on the basis that a test taker will leave an item blank if they don't know the answer, rather than risk losing a point for choosing a distractor. However, I don't know of any instance where this has been practised. This may be because it would introduce a construct-irrelevant factor in the form of the test taker's attitude to risk; test takers who are temperamentally cautious might be deterred from answering when in fact they do know the correct option.

Will multiple-choice tasks not always correlate with other cognitive skills up to some point?
If you mean with non-linguistic cognitive skills such as problem-solving (so test takers who are good at theses skills are favoured independently of their language ability), then yes, inevitably up to a point, which is why it is good to minimise those aspects of MC items.

If you mean with language competence in general, the term “multiple-choice” denotes a method, not a construct. I worked in a language school where we had a placement test consisting of 100 MC items targeting grammar, reading and listening (and a separate very brief interview to assess speaking). It worked pretty well as a way of placing students into groups by level, suggesting that the scores did indeed correlate with language ability up to some point, but I think that was because it sampled effectively across three domains of language use rather than because of the method. The method was used because it enabled us to test a large intake of students in the morning and place them in classes by lunchtime. Given more time, a mixture of methods – some MC, some other item types – would have been preferable.

How good are multiple-choice questions or selected response questions in general in assessing grammar?
Pretty good at assessing grammatical *knowledge*, i.e. the ability to apply grammatical rules in particular contexts when presented with these on paper. This is not the same as the ability to produce grammatical speech and writing in real time, of course. Purpura (2004) is good on this.

In your opinion, will there be any way to make a multiple-choice test that really gauges a person's true language abilities?
No. But “a person’s true language abilities” is an immensely complex set of interrelated constructs that no test can really measure. MC items have their uses, and their limitations. Even the best devised MC language test will provide no direct evidence of spoken or written production, for example.

What do you think of distractors like "all of the above" / "none of the above" or "not given" in a true/false question?
Haladyna (2004) advises (cautiously) against “all of the above” and “none of the above” on the grounds that they are vulnerable to test-wise strategies. “All of the above”, in particular, is weak as a distractor as the test taker can eliminate it if they recognise that any *one* of the other distractors is false.

Regarding “not given”, it can be very difficult to distinguish between *false* (the text says something to the contrary) and *not given* (the text says something else, but not this), so it is difficult to formulate items where this distinction is clear-cut. In reviewing items of this type, I have sometimes come to a different conclusion from the item writer. I would certainly advise strongly against this practice for listening.

References

Frary, R. B. (1995). More multiple-choice item writing do’s and don’ts. *Practical Assessment Research & Evaluation*, *4*(11), 1–3. Accessed 14 May 2020 https://scholarworks.umass.edu/pare/vol4/iss1/11/

Fulcher, G. (2014). Multiple-choice items. Podcast accessed 14 May 2020 http://languagetesting.info/features/mc/items.html
(google “languagetesting.info”)

Grosse, M. E., & Wright, B. D. (1985). Validity and reliability of true-false tests. *Educational and Psychological Measurement*, *45*(1), 1–13. Accessed 14 May 2020 https://doi.org/10.1177/0013164485451001

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah, NJ: Lawrence Erlbaum.

Holzknecht, F., McCray, G., Eberharter, K., Kremmel, B., Spiby, R. & Dunlea, J. (forthcoming). The effect of response order on candidate viewing behaviour and item difficulty in a multiple-choice listening test. *Language Testing*.

Purpura, J. E. (2004). *Assessing grammar*. Cambridge: Cambridge University Press.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13. Accessed 14 May 2020 https://doi.org/10.1111/j.1745-3984.2003.tb01102.x

Glyn Jones, 14 May 2020
https://cefrreplication.jimdofree.com