Reviewer’s report

Title: Adding to the debate on the numbers of options for MCQs: The case for not being limited to MCQs with three, four or five options

Version: 0 Date: 23 Jun 2019

Reviewer: Lambert Schuwirth

Reviewer’s report:

Comments concerning manuscript MEED-D-19-00513, entitled: ”Adding to the debate on numbers of options for MCQs: the case for having a variable longer list of MCQ options”

thank you for giving me the opportunity to review this manuscript. The author and I share a passion for these issues and it is always a pleasure to read such work.

I am a strong advocate for having variable numbers of options in multiple-choice questions and even for having multiple different question formats within one test. Actually, part of the research for my own PhD - unfortunately apart I never published - looked at the psychometric consequences of having a variety of question formats within one test using both a stratified/straight alpha comparison and a multivariate/univariate generalisability study. In that research we found that there is absolutely no factor associated with question format or number of options within multiple-choice. This simply corroborates with the literature that the author cites. So, he and I do not disagree at all about the suggestion to use variable options in multiple-choice questions.

I do have some suggestions and some concerns about the current manuscript though.

The opening sentence is an important part of the foundation for the argument and therefore would be better if it were referenced, for instance with "Rodriguez M. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educational Measurement: Issues and Practice 2005;24(2):3-13.

The argumentation as a whole is not put into a clear frame. The arguments to reduce or limit the number of options per multiple-choice are clearly made from a psychometric framework with a focus on discriminating between the more and less knowledgeable/competent students. The arguments for an increase/variability in numbers of options are clearly made from a more educational point of view. Part of them are also made from a content validity/direct validity (cf. Ebel RL. The practical validation of tests of ability. Educational Measurement: Issues and Practice 1983;2(2):7 - 10.) point of view. I think it would dramatically strengthen the argument is that different frames would have been made explicit.

There might be differences of opinion but the praeterio in lines 36 to 38 does not work for me. In such a praeteritio a suggestion is made implicitly for an argument which is then withdrawn immediately, the typical opening being "I am not suggesting that…, but…” The same is done
here. First, the suggestion is made that each of these studies are underpowered but than it is concluded that this conclusion is unlikely. For me this weakens the next arguments. If a study is unable to detect differences in psychometric parameters at the level at which decisions about students are typically made it is sufficient to claim that there are no relevant differences. That's not under powering.

The sentence in lines 45 and 46 does not make sense "can are not limited". I assume it was meant to read "can not be limited".

Starting a discussion about multiple-choice items with a different item type, extend matching items may not be most logical. According to the originators of extended matching items, Swanson and Case, the reason for the large number of options was to counteract cueing. This comes at the expense of variability of domain coverage, but given the large debate in the literature about the suitability of closed questions to assess clinical decision-making/reasoning and in the context of national certification/licensing it is a logical design decision.

To suggest that the options list could be published to students does not provide an argument as to why it should. The assumption that it will guide learning and clinical reasoning is not sufficiently founded in the manuscript, either by a clear rationale nor by supporting literature. For example, no argument has been put forward as to why this would not lead to more strategic revising and cramming for the test rather than support a more general 'studying to become a better doctor' - approach. To simply assume that this will automatically remind the examinee of relationships between diagnosis and symptoms and the realisation to need to differentiate between possible diagnosis and using the information, is too big of an inference in my opinion.

Longer lists of options, especially the long menus, have been suggested both by Veloski and me as an alternative to short answer questions and not as a variation on multiple-choice questions. They were both suggested as a type of easily scoreable open-ended questions. The reason for it being again the notion of cueing effect. This is why Veloski's paper calls it the un-Q-format. Both Veloski's and my approach were focused on designing an assessment format which would disallow the students to scan all the options before making a choice. That was central in both studies, and in my work I tried to demonstrate - perhaps a bit clumsily - that they behaved, psychometrically, more as open-ended questions than as multiple-choice questions. Therefore, I'm afraid that I think that they're not a strong component of an argument for allowing for a variable number of options.

The final paragraph of this section makes a lot of sense to me, but from an educational point of view. There is also literature supporting this suggestion, namely the ream of papers showing that the content of the question is more important with respect to its validity than the question format. If the content is overriding it should also be essential in determining the number of options. This is a direct validity argument, but it's certainly supported by literature.

I'm afraid I don't understand what is meant with the first sentence in line 66. This section makes a very plausible argument to me, again the content of the question should determine the number of realistic options (in our own work we've always advocated to first determine the content of the question, then determine how many realistic options that were before deciding whether an open-
ended question or a closed type question was best). For me, and perhaps also for the general reader, examples work very well. I would suggest to include them in a more elaborate way - writing them out completely - and make them stand out layout-wise.

The argument in lines 85 and 86 needs further explanation. Just saying that the assumption is erroneous and there is significant information begs the question what kind of information this is and how we know that this is significant information. For example, the notion of "clinically important distractors". This assumes that we all agree on what is a clinically important distractor and thus that we would have full agreement about relevance of questions and distractors. The reader may assume, which may be not entirely unfounded, that there is no such agreement. Therefore, it would be good if the manuscript were to contain some more information about how agreement was reached on what are clinically important distractors or which literature is there to support it. This is even more important if the inference from choosing an incorrect options, which is deemed unsafe practice, would actually mean that the student would also choose that option in a real practical patient case. That is an inference which is quite big and has often been argued against in the discussion about so-called killer stations in OSCEs. If it's not an inference we like to entertain with respect to clinical skills stations in OSCEs, why what we want to entertain it in multiple-choice questions?

The cohort level misconception in line 89 is probably an important aspect, but it assumes that 'clinically important distractors' are deliberately and purposefully included in the items to detect issues concerning curriculum development. If they are not, there is the likelihood that it's going to be a bit of hit-and-miss approach.

The argument in the sentence in lines 106, 107 and 108 is something I don't agree with. I am a big fan of looking at content of assessment, and in programmatic assessment triangulating of information on similar content across assessment methods, but when we are talking about multiple-choice tests the purpose is typically to produce a total score as a meaningful indicator for knowledge/competence. Even the most modern validity theories - including Kane - see the combination of items as leading to a score, universe score, target domain and construct. Part of that series of inferences relates to the connection between observed score and universe score. That is an inference which is not made on the basis of an individual item but how all items contribute to a total score. The improved reliability found by reducing the number of options is therefore not an artefact of the concurrent increased number of questions, it is the effect were trying to achieve. Quite the contrary, there is always the issue of domain specificity and it is more than likely that increasing the number of options per item and thus decreasing the number of items per hour of testing time plays into the hands of domain specificity and has psychometric consequences. These, in our current testing paradigms are not artefacts but true effects. So, even in the most modern views of validity that universe generalisation is a central link without which the whole argument is broken. One could argue that there are different ways to achieve that universe generalisation, but these would be more pertinent to different assessment methods the multiple-choice tests.

I realise very well that I've been perhaps very critical and I do apologise for that. But the reason why I am so critical is not because I disagree with a central tenet as I think makes much more sense to be flexible with a number of options in multiple-choice questions. That is why I think
It's important to put forward the best possible arguments, and I think these are not from a psychometric point of view but from a direct validity, educational value and educational consequence point of view.

**Are the methods appropriate and well described?**
If not, please specify what is required in your comments to the authors.

No

**Does the work include the necessary controls?**
If not, please specify which controls are required in your comments to the authors.

Unable to assess

**Are the conclusions drawn adequately supported by the data shown?**
If not, please explain in your comments to the authors.

No

**Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?**
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

Not relevant to this manuscript

**Quality of written English**
Please indicate the quality of language in the manuscript:

Needs some language corrections before being published

**Declaration of competing interests**
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?
4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?
5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I have no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal