Author’s response to reviews

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

Authors:

Mike Tweed (mike.tweed@otago.ac.nz)

Version: 1 Date: 25 Jul 2019

Author’s response to reviews:

I would like to thank both the reviewers and the editor for their comments. Given both reviewers agreed with the philosophy, but not necessarily the articulation, of the argument, I felt encouraged and have taken their comments and revised the article.

1.1 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 opening sentence has been developed and this reference has been added to the opening sentence.

1.2 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.

The psychometric perspectives being the argument for reducing the number of distractors and the clinical authenticity (linked to content) perspective being the argument of having a variable (potentially increased) number of options has been introduced in the background and woven within the argument.

1.3 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 In such a praeteritio a suggestion is made implicitly for an argument which is then withdrawn immediately, the typical opening being "I am not suggesting that…, but…” The same is done here.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.

These qualifiers to the argument for reducing and having a fixed number of options has been removed.

In addition the title has been re-worded to show that this is mainly a counter argument to having a small fixed number of options. The pro-argument has been referenced, now further including the reference suggested 1.1.

1.4 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". Changed.

1.5 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.

This section has been re-ordered and re-framed. Long list menu items are covered first and EMQ covered second, using the argument that although developed for different purposes, the purpose of clinical authenticity of multiple options is demonstrated.

The section on guiding learning is hypothesis and has been removed.

1.6 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.

This is now more prominent at the start of this section. As for 1.5, the point to be made is that longer lists developed to prevent cuing, also demonstrate how the number of options open to clinicians can be matched in longer lists.

1.7 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.

The importance of content and numbers of options is re-iterated (and referenced) in the closing paragraph of this section.

1.8 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.

An example has been written out in full, and hopefully will help potential readers see the issue.

1.9 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.

This opinion piece is not necessarily the place to explore how clinically important potentially unsafe have been defined in a variety of research settings. I have now included a sentence that this has generally been done using a panel of experts, but the way in which each panel undertook the task did vary.
It is true that the level of agreement between judges relating the unsafeness of potential responses does vary, hence the importance of a sufficiently sized panel of committed knowledgeable individuals.

1.10 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 important of unsafe responses not being equivalent to “killer” stations is included. A pattern demonstrated form a program of assessment, with several unsafe responses, persisting despite feedback and the opportunity to improve would be a concerning pattern.

1.11 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 articles referenced did not purposefully ensure that every question had an equal number of “unsafe” answers. In fact the number of unsafe options should vary with the content. So yes, there is the potential for missing a cohort level misconception not on the option list.

1.12 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.

This section has been removed, as the argument was weak and controversial as demonstrated by the comments of both reviewers.
1.13 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.

I thank the reviewer for their frank and helpful comments. I believe that the changes made in light of the comments have improved the argument.

2.1 Let me begin by stressing that I do not disagree with the author's overall contention. I think the existing literature supports the notion that content relevance should be paramount when deciding on how many MCQ distractors to use for a specific item. More generally, the content of items are much more important than their format.

This importance of content has been included, as per 1.2 and 1.7.

2.2 However, there are many factors to consider when determining the number of options of MCQ distractors and the paper in its current form does not adequately deal with most of these. Finally, the overall position seems overblown - I think the case for having a variable longer list of options does exist, but only in certain circumstances (which require more nuanced articulation), and only when supported with specific content rationale and psychometric evidence.

The argument has been changed as suggested by review 1, such that clinical authenticity (content) and education perspective are the main arguments for a variable likely longer list of options, differing across questions.

The conclusion has been toned down to “consider a move to”

2.3 No evidence is given throughout to substantiate the arguments. In fact, they are more assertions than arguments. The overall thesis that assessment organisers should consider having variable numbers of options is not controversial. Having an increased number of options for some specific items, if justified, is also not controversial. However, the justifications given are not without counterargument. If the author wishes to pursue the paper further, it would be good to at least anticipate the counterarguments to the position given, and to make some comment on them.

I agree with the reviewer, but I think there are institutions who have taken the evidence on reducing the number of options to 3 or 4, and are looking to institute this as policy such that all questions have the same number of options (be it 3 or 4).

As per preview 1 comments (1.2), the argument has been changes to the differing perspectives.
2.4 I think the biggest shortcoming is the failure to appreciate psychometric perspectives, especially based in Item-Response Theory or Rasch Measurement. No evidence based on validated models of MCQ assessment is provided.

I have changed the title to reflect that this is the argument for a variable likely longer number of options, and that the arguments for a shorter fixed numbers of options is as per the literature, including the references suggested by reviewer 1.

2.5 The main thread running through the text involves a range of tautologies, such as there is no reason why the number of options has to be the same; and it is feasible to deliver tests via online platforms with different numbers of options. Neither of these statements necessarily mean that longer lists of options are best.

I have moved practicalities to the final paragraph. This was an attempt to meet the counterarguments of practicality (that I have heard).

2.6 "Decisions related to broad clinical scenarios can are not (sic) limited to a small number of options"

This is probably the strongest suggestion in the paper as it relates to content relevance. Item quality is actually the most important element here, but, where relevant, more distractors may play a role. This does not mean, however, that more distractors should be used. Can does not imply ought. I think this assertion also falls over when considering the process of assessment through an item. The item is an indicator of competence, tied to the purpose of the exam, the articulation of a construct, and in the context of a blueprint to a curriculum. The item should be developed to ensure minimal false-positive and false-negative responses from candidates. In terms of measurement, more distractors add noise to the process, and compromise the 'purity' of the measurement process. They also run the risk of introducing a range of factors that are not content related, and thus are subjected to test-taking strategies of candidates, often resulting in a reduction in item discrimination.

The typo has been corrected.

I agree that purpose is important, and have included assessment purpose to guide learning and subsequent performance and also to make progression decisions. I think this is more important in the section on potentially unsafe responses and have included it there.

I would disagree that more distractors increase the noise (if noise is bad). If the clinical scenario has 10 options then why not include all ten?

The purpose of the assessment should be to guide learning and potentially inform progression decisions.

I would also suggest, if reducing the balanced elements from 8 to 3,4 or 5, the decision as to which ones to remove might introduce noise. (see below)
2.7 "Options lists should include all possible combinations of factors"

I don't agree with the assertion that they 'should'. They certainly can, but again this can seriously compromise the functioning of the item. Also, 'all factors' is a subjective thing when it comes to medicine. When drafting an item I would start with the 'most relevant' factors and then cut out any that are superfluous and likely to add noise to the assessment data. Still, the position given simply strengthens the notion of having variable numbers of distractors. It does not necessitate long lists of them.

I have now included an example in full, as suggested by reviewer 1.

A good point is raised about most relevant distractor options. Most relevant defined how. Distractors chosen by nearly satisfactory examinees, very poor performing examinees, or very rarely chosen (but clinically very relevant)? I would suggest all these, as in the next section.

2.8 "Options that are rarely chosen can provide information regarding students and/or for students"

"Many of the analyses upon which the recommendations to reduce the number of options are based on the assumption that all incorrect responses are equivalent". Do they really? I would be very surprised if this was the case. Distractors work different for different questions. Anyone involved in the analysis of MCQ papers using IRT/Rasch appreciate this. But distractors do become irrelevant when they are a 'waste of space' - practically no candidates go for them. This just adds unnecessary reading loads onto time poor test-takers.

I have qualified this to differentiate that there is a difference between psychometric performance of a distractor and clinical importance of a distractor is that was a response in authentic practice.

2.9 If one candidate selections a dangerous incorrect response, this may be useful information for students and teachers. But the outcome may also be because they ran out of time and guessed. It may also be because the question was poorly worded. While diagnostic feedback is incredibly important and distractor options should inform feedback processes, including long lists will not necessarily add value.

I agree, but guessing and making an unsafe choice is not good clinical practice.

A single unsafe response is unlikely to be used apart from feedback. A pattern of repeated unsafe responses despite feedback and learning opportunities must be a cause for concern.

A long list may be needed to include unsafe responses that are important for the content.

2.10 "Score reliability can increase with numbers of options"

I find this the weakest conjecture of all. It can, but that doesn't mean it will. The most important factors that influence reliability are item discrimination and item fit (when using IRT/Rasch
models). Not to mention just running more items. I worry that this proposition is based on a simplistic, classical approach to MCQ item analysis.

This section has been removed, as the argument was not well founded.

2.11 "Longer lists of options are feasible for examinees to use, especially given computer based".

I have no issue with this. Yes it is feasible. That doesn't support the author's contention that is should be done. As a final point on feasibility though, this is fine.

I find the conclusion weak and simply another conjecture. Also, the final part of it is non-controversial. The paper requires more detail, evidence and needs to be buttressed with a response to potential counterarguments.

As noted above this is included only as a counterargument to feasibility. It is now presented as such.

2.12 A few minor typos:

- in the abstract, page 2 line 19 "can are not limited to"

- page 3 line 45.

Corrected.

I have included a tracked changes version as supplementary material.