Reviewer’s report

Title: Student progress decision-making in programmatic assessment: can we extrapolate from clinical decision-making and jury decision-making?

Version: 0 Date: 31 Jan 2019

Reviewer: Wolf Hautz

Reviewer's report:

Dear authors,

thank you for the opportunity to review your manuscript on the comparison of decision making in programmatic assessment with clinical decision making and jury decision making.

I read the paper with great interest and applaud you in particular for looking into an example of group decision making well outside medicine, i.e. jury decisions.

To further improve the paper, I have several suggestions. My key concern is that some of the claims made in the manuscript appear rather bold given the available evidence. I would suggest to provide a more balanced argument. For example, while you identify a number of similarities between decision making in assessment and decision making in clinical practice, there are some important differences between the two that should be considered and that - in my humble opinion - limit transferability of findings from one to the other. While assessment decision making is a pretty well defined task, where - at least in theory - more information is always available, the information quality can be assessed and improved and time presser is limited, much of clinical decision making happens in rather ill-defined environments, where information is limited, of questionable quality and time pressure is high. I would suggest that in addition to drawing parallels between the two, to also elaborate on the limits of this analogy.

The same applies to your discussion of bias. While you prominently cite Crosskerry as a proponend of the importance of bias and debasing strategies, others (Norman most notably, Gigerenzer outside medicine) have heavily questioned the relevance of bias to real world decision making. Except for a small effect in only the most difficult cases that Mamede found for deliberate reflection, there is very little experimental evidence that generic debiasing strategies actually improve decision accuracy at all. Norman suggested, that far more important in the determinanption of decision accuracy than any bias is content relevant knowledge. I am not suggesting the issue is solved yet; I am just suggesting that your manuscript should present both points of views, because of the consequences for what you label "good practice".

I would also suggest to rename this section and be more modest here. Good practice implies that the available evidence is strong enough to identify such practice. While your manuscript does a great job of identifying relevant issues in assessment decision making and potentially relevant evidence from other fields, I would suggest that actual evidence in programmatic assessment is to date rather limited and transferability from other fields has not been tested.
A last example in need of a more balanced and nuanced discussion is your call for students to be considered incompetent until proven otherwise. While this idea has some intuitive appeal, the single reason you provide for this (rather consequential) practice is that medical schools have an obligation towards society to ensure physician competence and thus - if in doubt - should better fail candidates falsely than passing them falsely. However, one could argue that academia also has the responsibility not to waste societal resources by falsely failing students in which a considerable amount of educational resources has been invested by then. Furthermore, while falsely failing students where there is an unlimited supply of others may be defensible, many nations actually have a shortage of physicians and academia has the obligation to provide sufficient staff to meet society's demands. I am not suggesting it's better to falsely pass students, but I would raise the question whether a borderline competent physician is not better than no physician at all? Briefly put, this issue also needs a more balanced discussion and is far from decided upon yet. It should thus not be labeled a "good practice".

In addition, Gigerenzer and Todd have repeatedly shown that humans tend to stick with the default. Their catchy example: in some south german towns, subscription to "green energy" is around 90%, whereas in neighboring towns, its only around 10%, without any obvious explanation for this difference. In fact, when inhabitants in these towns voted on either renewable or traditional energy, opinions were split around 50/50 in all of them. But: towns, where a slight majority voted for renewable energy made this the default, whereas traditional energy became the default in the others. People are free to deviate from the default in all towns, but this requires thought, deliberation and action. As a consequence, people tend to stick with the default. Consequently, setting a default such as failing students should be very well justified.

A last issue I would like to raise is a conceptual problem with programmatic assessment per se. Integrating results from several assessments could be done in multiple ways. One could require students to pass each assessment separately, or one could average over several assessment's, so that good performance on some cancels out bad performance on others. The first one implies that each topic assessed is sufficiently important to require some degree of mastery from every student. The second suggests that in order to be labeled competent, one simply needs to be good at something, which can then make up for being not so good elsewhere (which, I would argue, reflects clinical reality with its many specialties quite well). A third way would be a combination of the two approaches, where some topics are labeled sine qua non, i.e. must be passed, and others are averaged over. This however gives different weight to different topics - something I am not sure many faculty boards can agree upon. The resulting underlying question here is: is there such a thing as a competent physician? Given that content and context specificity is the "one truth in medical education", I am not sure competence can be judged without respect to context and content. I, myself, consider myself competent on X,Y and Z, but certainly not on A, B or C, also all those could be medical subjects. I would be interested in how programatic assessment addresses this problem.

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

Acceptable

**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.

none

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