Author’s response to reviews

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

Authors:

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

Timothy Wilkinson (tim.wilkinson@otago.ac.nz)

Version: 1 Date: 26 Mar 2019

Author’s response to reviews:

We thank the editor and reviewers for their very helpful feedback on our paper. This has prompted a fairly substantial revision of our paper, in ways that we believe have improved it considerably. We provide issue by issue responses to the comments, as below.

Firstly in light of comments we have changed the title to better reflect the content.

In light of comments we have changed the title to better reflect the content.

1. The Abstract should be structured with a background, main body of the abstract and short conclusion.

   This has been done

2. Please note, Reviews should be structured so that they contain a "Background" section, please include a Background heading in your manuscript.

   This has been done

3. Please amend the Availability of data and materials statement to: "All data generated or analysed during this study are included in this published article".

   This has been done
Thank you for the opportunity to review this manuscript. The topic is both of great current interest to the medical education community as well as being the subject of my doctoral and ongoing research (Interdisciplinary approaches to risk and uncertainty, especially as they apply to professional judgements and decision making) and educational work (e.g. qualitative and quantitative data synthesis and integration in qualitative, mixed methods and multimethod research) Detailed comments as well as overall comments have been made in the attached annotated manuscript. One of the key strengths is the exploration of ideas from fields other than usually cited in medical education and the extensive number of citations listed. We thank the reviewer for their suggestion. No response required for this.

As an invited commentary, methodological rigour may be less important than for an empirical study. However, I would like liked to have seen some critical appraisal of the literature and research cited so that the limitations, as well as more reference to the "classic" or primary literature from the disciplines being cited. For example social psychology literature has cited to support analogy with decision making by juries, rather than from law. We note that as an invited commentary, rather than a systematic review, there is allowance made for the lack of critical review of all the literature. We have reviewed claims made that are based on the literature and lessened some of other claims. We have cited reviews in the interest of limiting the number of citations, and in terms of summarising several manuscripts focusing on relevant topics. Most of the literature related to group decision making by Jurors and juries has been developed by and published by social psychology rather than those undertaking academic study of law.

Another example is heuristics and biases from selected medical literature rather than from the originators of the field such as Tversky, Kahneman, Slovic etc (all psychologists). Some of the analogies could be better supported by stronger arguments.

As we focus on the clinical application of this primary research (quoted by the reviewer), much of this is covered by Elstein, Croskerry (cited). We intended to keep this related to clinical decision making, given the readership is expected to be those familiar with clinical-decision-making.

Recommend also making it clearer at the outset the the range of decisions in programmatic assessment and that this review only covers high stakes progression decisions and how these are often made in medical schools. Suggestions for these are included in the comments.
This has been done.

In summary it could be a useful addition to the ongoing debates about assessment decision making, but could more clearly lay the foundations for further work to be done.

We accept that there is considerable debate about clinical-decision-making and its psychological underpinning and application. There are those that believe the role of heuristics and biases is overestimated compared to underpinning knowledge. We chose not to extend the scope of this paper into this complex area.

In some ways it could have been a more satisfying read if only one analogy (eg clinical decisions, heuristics and biases and lessons from the extensive literature there OR legal decisions and judgments) had been made and explored in greater depth with more critical analysis. Thank you again for the opportunity to review and make suggestions for improvement.

We have kept both clinical and jury decision making as we feel they complement each other and support the tips.

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.

We thank the reviewer for this comment and feel no other response needed to this comment.

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.

We have included the limitation of partial information and time pressure for clinical-decision-making. Whilst we accept clinical decision making is not a perfect analogy, it is useful as it will be familiar to the readership. We admit we have been selective in drawing analogies in both clinical and jury decision making, but we do raise the limitations of both analogies respectively.

The same applies to your discussion of bias. While you prominently cite Crosskerry as a proponent 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 determinant 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".

As in response to the prior reviewer, we accept that there is a considerable literature questioning the role of heuristics and bias in clinical decision making, and therefore the role in being aware of these biases, as opposed to increasing underpinning knowledge. As an opinion piece rather than a systematic review, we believe that introducing this debate, might then require a debate on robust progression making. We chose not to extend the scope of our paper into this area.

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.
We have reduced this claim from good practice to suggestions. As an opinion piece we believe that it is reasonable to look for analogy with some limited transferability.

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". We have included the counter argument of duty of care to students, but have not gone as far as the idea that substandard doctors is better than no doctor in some circumstances. This could be a whole complete article as a matter of debate.

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.

We have included theory-perseverance effect related to clinical decision making. Whilst the German towns example is interesting, we feel it does not fit in with our analogies of clinical decision making or jury decision making.

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.

We have not included a detailed review of programmatic assessment. The review of deconstructing programmatic assessment, does include aggregating information over time, which is not necessarily the same as independent hurdles or average. Going beyond this is beyond the scope of this opinion piece.

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.

The question of true competence is interesting, but beyond the scope of this piece. Again this could be an article as a debate stand-alone.