Author's response to reviews

Title: A randomised trial of the influence of racial stereotype bias on examiners' scores, feedback, and recollections in undergraduate clinical exams

Authors:

Peter Yeates (p.yeates@keele.ac.uk)
Katherine Woolf (k.woolf@ucl.ac.uk)
Emyr Benbow (emyr.benbow@manchester.ac.uk)
Ben Davies (bgjdavies@doctors.org.uk)
Mairhead Boohan (m.boohan@qub.ac.uk)
Kevin Eva (kevin.eva@ubc.ca)

Version: 2 Date: 21 Aug 2017

Author’s response to reviews:

Dear Editor,

Thank you very much for offering us a repeat opportunity to re-submit our manuscript, and to respond to the reviewers additional comments. We have reflected on each of them, and made some further revisions to the manuscript whilst offering our reasons for keeping some elements unchanged. We believe that this has further strengthened the manuscript, and we’re grateful to the reviewers for their insights. We believe that the study continues to be of importance and will be of considerable interest to your readers.

Please find below a point by point response to the reviewers comments, indicating where changes have been made.

We would be more than happy to address any further comments or queries which you may have. Thank you very much for your ongoing consideration.

Yours sincerely,

Peter Yeates

Reviewer reports:
Reviewer #1: Thank you for responding to all of the comments made regarding revisions to the manuscript and incorporating them into this revised version. I believe that the clarity of the writing will make the paper valuable to readers. Since the findings differed from expectations, I believe it will make it clear that additional research regarding examiner bias is necessary. As it is, the paper provides evidence that supports the validity of the process. It does not appear that examinee or examiner characteristics introduce construct-irrelevant variance to the ratings. Though the data are based on a simulation, they provide some insight to how stereotypes could be activated and then affect examiners' objectivity in rating performance.

***Thank you for your kind comments. We are glad our revisions met your satisfaction.

Reviewer #3: In my opinion, the authors have done a great job in responding to my own and the other reviewer's comments.

***That is great news. Thank you.

I have a few comments:

- With respect to my comment on contextual differences: overall I am satisfied with the explanation added by the authors (lines 5-12). However, research has shown - at least in the job application context - that more experienced raters tend to rely more on intuitive judgments and hence might be more at risk for bias rather than less compared to less experienced raters. Therefore, I'm not sure whether more experience per se might lead to less stereotyping (although a tendency to individuate might do so).

***We are glad the reviewer is satisfied with these revisions. The original review raised the question of why the medical education context might differ from the context of job applications, which is the issue we addressed on page 12 (lines 5-12) by indicating that while unexplainably different, our results align with other observational studies and, thus, appear robust. The notion that more experienced raters might respond differently than less experienced raters was not raised in the original review. We have noted, however, at the top of page 9 and in table 2, that both groups were similar in all measured demographics. As a result, if more experienced examiners showed greater bias, as hypothesized by this reviewer, then the null results we have reported could only be derived if less experienced examiners showed greater bias in the opposite direction. Given that we know of no theoretical reason for that to be true, know of no findings suggestive of such an interaction, and given that the experimental design was not powered in a way to allow exploratory post-hoc tests of differential effects in every subgroup that could be generated by the demographic information collected we believe it would be a mistake to speculate about this possibility and have not changed the manuscript for fear of lengthening the text in a manner that distracts from the central and answerable questions. We can of course add a sentence of limitations in this regard if the Editor thinks it crucial.

- With respect to my comment on the reliability of the mark sheet: I was referring to inter-rater reliability - the degree to which different judges or raters agree in their assessment
decisions. My comment was invoked by the data presented in the table at the bottom of page 24, which suggest that the panel of 8 reviewers rated the videos as less similar with respect to communication than with respect to knowledge. This finding would be in line with other studies that have suggested that the inter-rater reliability for communication scores is (often) lower than for knowledge scores. So, I was wondering whether the fact that raters tend to disagree more on scores for communication could somehow have influenced the findings of the current study.

***Yes, we assumed this comment to be about inter-rater reliability, but the data on the bottom of page 24 indicate only that the range of communication scores was greater than that of knowledge scores, not that the communication scores were less reliable. The variance in scores is made up of variance attributable to the students in the videos, the raters who rated them, and a residual error term (the student x rater interaction). The reviewer assumes that the larger range in communication scores is due to greater variation across raters (which would suggest lower reliability), but in fact it could just as easily arise because the students themselves were more differentiable based on their communication scores than they were based on their knowledge scores, which would suggest greater reliability. In other words, we do not know if raters disagreed more on communication scores because one can make no claims about reliability based on the range of scores alone. As we indicated in our first response, reliability analyses are not appropriate or relevant to experimental designs. Even if they were to be calculated, however, there is no reason why this would mitigate a systematic effect of student ethnicity on scores. As a result, we again, do not think it appropriate to alter the manuscript in response to this comment.

- With respect to my comment on the test of stereotype activation: I am very happy with the addition of this issue to the limitation section. I also agree with the authors that this does probably not explain the lack of an effect of stereotype activation on scores or feedback. However, I still think that it might explain the lack of differences between the two groups. The authors hypothesized that Group A would have faster response times to stereotypical words than Group B, but this hypothesis was not confirmed. If the stereotypes are already active before completing the task, it might be that the exposure to three student performances is not enough to make a difference between the two groups.

***Thank you again for recognizing the value of our revisions. We would summarize this discussion as follows: Stereotype activation might occur through general (non-stimulus specific) activation (which we did not control for) or stimulus-induced activation (which we tried to manipulate). The fact that stereotype activation was observed, but without difference between experimental groups suggests that the lack of difference in ratings arose due to general activation, not stimulus-specific activation. We interpret the reviewers’ challenge, therefore, as suggesting that perhaps stimulus-specific activation (i.e., activation caused by the observations made of the students themselves) is required to see the influence of stereotype activation on raters’ judgments of the students. To acknowledge that issue we have altered the final sentence in the limitations on page 12 and added one additional sentence such that the end of the limitations reads as follows: “Lastly, as we did not have a control condition in which participants performed the LDT without watching the videos, we cannot definitively claim that participants activated an Asian stereotype specifically in response to the videos rather than that the activation observed was already present upon beginning the study. These results support the central conclusion that examiners had a stereotype which was active at the time of judging
performances, which does not appear to have been applied to their judgements. However, it is possible that stereotype related differences in judgment are induced only when the behaviours of the individuals being judged contribute directly to further stereotype activation.”

- With respect to my comment on the description of the results on the influence on examiners' memories: to me this section is clear now.

***Excellent.

- With respect to my remark regarding BAPIO vs RCGP and GMC: although the sentence is much clearer now, I would still suggest the authors to explain the acronyms for non-UK readers.

***On re-reading this sentence, it has become clear that specification of these acronyms is actually peripheral and largely irrelevant to the point made. Rather than adding an extensive amount of text to clarify the parties involved in the legal case for non-UK readers who would also require a description of each organization we have simply deleted the acronyms such that the sentence now reads more simply: “A recent prominent legal case in the UK (see [26]) has reaffirmed an important principle of equality law”

- As an additional question: what happened with the data of the participants who guessed the study's true purpose? It appears as if no action was undertaken. Would it not have been better to exclude their data?

***This is an important issue in principle, which is in fact why we made an attempt to determine if participants recognized the study’s true purpose. In reality, however, given that 2 people fell into this category out of 159 participants (1.2% of all who completed the study) it is both impossible to treat them as a sufficiently powered subgroup and inconceivable that their inclusion would have an influence on the robust patterns reported. As a result, there is nothing to be gained from the time that would be required to re-do this analysis and we have simply responded by noting explicitly at the top of page 9 that all participants were retained: “Examination of participants’ responses in the debrief phase indicated that only 2 participants (1.2% of all completed respondents) guessed the study’s true purpose; therefore, no participants were excluded given that all consented to ongoing inclusion of their data.”

As a smaller remark: the authors may want to take another look at lines 31-34 at page 4, since both sentences start with "as a result”.

***We are particularly grateful to the reviewer for offering this “smaller remark” both because the awkwardness was actually worse than indicated as we used “as a result” three times in short succession and because re-reading that text alerted us to an inaccurate reference to “Halo effects”. “As a result” we have modified the text and adjusted the reference list to accommodate the deleted citation.