Reviewer’s report

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

Version: 0 Date: 23 Jun 2017

Reviewer: Karen Stegers-Jager

Reviewer’s report:

Overall evaluation and general comments

This was an interesting paper on examiner bias in undergraduate clinical exams. Surprisingly, the hypotheses regarding the existence of bias were not confirmed, despite the fact that these originated from a whole body of existing literature from other domains. Although the authors provide some explanations for these non-findings in the discussion, I still wonder why these consistently reported mechanisms would not apply to the context of undergraduate clinical exams. What makes this context different from for example that of job applications? And how did the authors ensure that the 'Asian' videos were experienced as typical Asian performances? It might be that in real practice communication skills for Asian students are rated as being poor for different reasons than those of white students, for example due to differences in communication styles.

Overall the paper is very well-written and clear. See below for some suggestions per section.

Title and abstract

The title is a good reflection of what was looked at in the paper, but it is a bit long. I wonder whether a shorter would attract more readers.

The abstract is very clear.

Introduction and Conceptual Framework

The introduction establishes a clear conceptual framework, identifies the gaps in the literature and concludes with several research questions. I’m only not sure about research question 3, about feedback. In the methods it turned out that authors are looking both at the focus and the valence of feedback, however, it is not clear from the introduction why they did this. Why do the expect the valence of the feedback to be different? Following the differences in performance I can understand that they expect differences in the amount of positive or negative feedback, but that is different from poor versus good feedback right? So why would the quality of the feedback be different?
Methodological Rigour

Overall the description of the methods is very clear, however, I do have some questions and remarks.

1) With respect to the videos: how many different actors were used? (I suppose 6?) And were the actors male or female?

2) With respect to performance scoring: was the mark sheet specifically designed for this study, or were the raters familiar with the mark sheet? What was the reliability of the mark sheet? (as suggested by the numbers in the Appendix, the rating of communication skills may have been less reliable?)

3) As described earlier, I have some problems understanding the concept of valence. After reading the appendix it became clearer, but I think the reader should be able to get a basic understanding without reading the appendix. The main problem, probably, is the use of the words poor and good feedback, which refers in my opinion to the quality of the feedback, where I think the authors intend to refer to positive and negative feedback, which is about the content of the feedback.

4) Regarding the test of stereotype activation, I wonder whether the authors can really claim that the stereotypes were activated by the judgment task, as they did not use a control group that did not perform the judgment task. Might it not be that these stereotypes are always present? And might this not be an explanation for the fact that they did not find any differences between the two groups?

5) At the end of the section on test of memory it would help to repeat the focus of each of the groups (stereotype consistent or inconsistent) in the hypothesis. Also in this section it is not clear whether the 20 accurate statements only came from the mixed performance. (sentence 38-39)

6) With respect to the test on recollection I wonder why an ANOVA was used instead of a t test as only two groups were compared.

Results

Overall the results are clearly presented. I have just two comments.

1) The statement that the activated stereotypes did not influence their provision of feedback made me wonder whether the level of stereotype activation could also have been included in the analysis?

2) The description of the results on the influence on examiners’ memories was not totally clear to me. For example, lines 19 to 22.
Discussion and Conclusions

The discussions section is fine. The conclusions are clearly stated and they relate back to the introduction. However, as indicated above I would suggest the authors to think a bit more on what makes this context different from others in order to explain their non-findings. Also I would like to read a bit more on the practical implications of their study.

As a smaller remark: I got totally lost in the sentence BAPIO vs RCGP and GMC.

Clarity, Length, and Ethical approval

The paper is well written and ethical issues have been addressed appropriately. Length of the paper is ok. No reasons for concern.

I suggest the authors to have another look at the Appendix section on validation of good and poor communication. Eg, they speak about 7 and 8 reviewers, and there are some problems with the sentences on Cohen’s Kappa.

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

Yes

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

Yes

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

Yes

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.

I am able to assess the statistics

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.

'I declare that 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/](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.