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: 1 Date: 23 Jul 2017

Author’s response to reviews:

Sunday 23rd July

Dear Editor,

Thank you very much for offering us the opportunity to re-submit our manuscript, and the opportunity to respond to the reviewers comments. We have reflected on each of them, and made several revisions to the manuscript as a result. 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. A short section at the end describes some small additional clarifications which, on reflection, we felt would further assist readers to clearly understand the work.

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
A randomised trial of the influence of racial stereotype bias on examiners' scores, feedback, and recollections of the performances of British Asian medical students in undergraduate clinical exams

Peter Yeates; Katherine Woolf; Emyr Benbow; Ben Davies; Mairhead Boohan; Kevin Eva

BMC Medicine

Dear Dr Yeates,

Your manuscript "A randomised trial of the influence of racial stereotype bias on examiners' scores, feedback, and recollections of the performances of British Asian medical students in undergraduate clinical exams" (BMED-D-17-00744) has been assessed by our reviewers. They have raised a number of points which we believe would improve the manuscript and may allow a revised version to be published in BMC Medicine.

Their reports are below. Please also take a moment to check our website at http://bmed.edmgr.com/l.asp?i=93972&l=RDUN2D7H. Please note that as BMC Medicine has a policy of open peer review, you will be able to see the names of the reviewers.

Additional revisions added at this stage are allowed at your discretion. Please make sure to add a section detailing these in your response letter.

If you are able to fully address these points, we would encourage you to submit a revised manuscript to BMC Medicine. Once you have made the necessary corrections, please submit online at:

http://bmed.edmgr.com/

If you have forgotten your username or password please use the "Send Login Details" link to get your login information. For security reasons, your password will be reset.

A point-by-point response letter must accompany your revised manuscript. This letter must provide a detailed response to each reviewer/editorial point raised, describing exactly what amendments have been made to the manuscript text and where these can be viewed (e.g. Methods section, line 12, page 5). Please also ensure that all changes to the manuscript are indicated in the text by highlighting or using track changes. If you disagree with any comments raised, please provide a detailed rebuttal to help explain and justify your decision.

Please also ensure that your revised manuscript conforms to the journal style, which can be found in the Instructions for Authors on the journal homepage.

A decision will be made once we have received your revised manuscript, which we expect by 26 Jul 2017.
Please note that you will not be able to add, remove, or change the order of authors once the editor has accepted your manuscript for publication. Any proposed changes to the authorship must be requested during peer-review, and adhere to our criteria for authorship as outlined in BioMed Central's policies. To request a change in authorship, please download the 'Request for change in authorship form' which can be found here - http://www.biomedcentral.com/about/editorialpolicies#authorship. Please note that incomplete forms will be rejected. Your request will be taken into consideration by the editor, and you will be advised whether any changes will be permitted. Please be aware that we may investigate, or ask your institute to investigate, any unauthorized attempts to change authorship or discrepancies in authorship between the submitted and revised versions of your manuscript.

I look forward to receiving your revised manuscript and please do not hesitate to contact us if you have any questions.

Best wishes,

Paulina Szyszka, PhD
Associate Editor, BMC Medicine
https://bmcmedicine.biomedcentral.com/

Reviewer reports:

Reviewer #1: This was a well-designed and well-executed study. It is valuable to conduct work in the accumulation of evidence of the validity of scores and decisions based on various assessment instruments. This work showed that although examiners may activate stereotypes of ethnic minorities, it does not appear to affect their scoring. Because of the this manuscript is so well written, I have only minor, discretionary edits to suggest.

*** Thank you for your positive review.

First, I thought that it would be ideal to report whether the study was approved (or exempted) by an ethics board. You detail the recruitment and consent of participants, including checking to see if their understanding of the study would lead to socially acceptable responses. Just a sentence noting review/exemption is needed.

*** The manuscript contains a statement about ethical review, indicating that ethical approval was granted by the University of Manchester ethics committee and lists the committee reference number. To fit with the journal style, this is in the declarations section, line 1-5 of page 14.

In the Methods section, you explain the scoring used in the scope of the study. While I am sure that the domains measured (knowledge and communications) overlap with those that examiners have rated in the past, I did wonder how the rating form compares to the task they complete in real life. It may be that the form that was used in the study allowed them to provide ratings that
reflected the simulated performances despite stereotype activation while those typically used do not.

*** Thank you for this suggestion. The mark sheet was derived from the standard format for OSCE score sheets from one medical school which recruited examiners. It was a typical domain-based mark sheet which is expected to have fairly broad familiarity to examiners. This has been clarified at lines 1-4 of page 6 of the manuscript.

Also, the lack of prior information on ratings from these examiners was managed by consideration of the standard deviation of the groups for power analysis (very smart decision). However, should the study be replicated, it could be that having this information provides an additional covariate for analysis of the scores the examiners provide in these simulated sessions.

*** We were slightly unclear as to the nature of the additional information requested at this point. As participants were anonymous, it was not possible to obtain prior information on ratings from examiners.

Again, congratulations on reporting on such well-designed work. Work in psychology has shown that there are several factors that can influence bias, and you have focused on a salient factor that could cause differential performance. I look forward to seeing this published.

*** Thank you again for a positive review

Reviewer #3: 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?

*** This is an important point, which we have considered further at lines 5-13 of page 12 in the discussion. Whilst we discuss a possible mechanism which has some potential to mediate the difference, we also note that further research is needed to understand the difference. Nonetheless, the findings in our study resonate with prior observational research in medical education which suggests a lack of ethnicity bias. As a result this counter-intuitive finding appears to be replicable.

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.

*** To script the videos, we followed the description of a stereotype of Asian students’ performance which has previously been published by Woolf et al. This relies on the discrepant
observation of good knowledge but poor communication skills. We validated the presence of these constructs in the scripts prior to filming. Additional description has been added of the scripting process and its validation. This is at lines 9-13 of page 5. In order to achieve experimental control, communication style was deliberately harmonised between the paired performances by the professional actors. The potential for a different sample of performance by Asian students to produce an effect is already acknowledged as a limitation of the study in lines 35-36 of page 12.

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.

*** We have reflected on the title and shortened it, whilst retaining the majority of meaning it conveys.

The abstract is very clear.

*** Thank you

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?

*** Thank you for this insightful point. To further explain the potential influence (and educational implications) of stereotype bias on feedback, a short additional section has been added to the background in lines 22-33 of page 3. Additionally, the 3rd research question has been amended to clarify the fact that it refers to both the valence and focus of feedback. The sentence leading to the hypothesis regarding valence of feedback has been enhance to reflect this theoretical grounding in lines 30-31 of page 6.

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?

*** Details of the actors have been clarified in lines 15-22 of page 5.

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?)

*** Thank you. The first question is a similar query to that raised by reviewer 1 and has been clarified in lines 1-4 of page 6 of the manuscript. Regarding the reliability of the mark sheet, we are unsure to what the reviewer is responding as we make no mention of reliability in the body of the text or the appendix. We presume it is the larger range noted for communication scores relative to knowledge scores, but that difference is not statistically meaningful and tells one nothing about the reliability of the scores. Reliability is a measure of how consistently one can discriminate between subjects of measurement (the individual performances in this case). It is irrelevant to experimental designs (as per Cronbach, 1957) because such discrimination requires considerable between student variability, which is considered ‘noise’ to the experimentalist whose purpose is to determine if a difference exists between groups that is large relative to the within group variability. As such, we think it would be inappropriate to confuse matters by making reference to the reliability of the scores collected using this mark sheet. While internal consistency (Cronbach’s alpha) can be assumed to be high between items within each domain examined, given that people tend to form holistic judgments, testing the reliability of the mark sheet was not the purpose of this study.

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.

*** Thank you for this point. The description of the content analysis has been moved to the main methods section (lines 8-23 of page 6), and references to good or poor feedback have been removed.

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?
*** Thank you for this point. Whilst Lexical Decision Tasks are a well-recognised measure, and are frequently used within experimental contexts as a measure of induced stereotype activation, we recognise, on reflection, that in the absence of a control group we cannot exclude the possibility of general background stereotype activation rather than induced activation. We have added a statement to the limitations section acknowledging this (lines 36 to 41 of page 12). Nonetheless, we do not believe that this explains the lack of an effect on scores or feedback. Regardless of whether stereotypes are a pervasive (rather than induced) phenomenon, they still have the potential to produce a cognitive biasing affect. As a result we stand by our conclusion that, despite being active (even if not activated) they were not applied to the judgement.

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.

*** The focus of each group (whether stereotype consistent or inconsistent) has been added in lines 5 and 7 of page 8 of the manuscript

Also in this section it is not clear whether the 20 accurate statements only came from the mixed performance. (sentence 38-39)

*** All 20 accurate statements did indeed come from the mixed performance. This has been clarified in line 29 of page 7 of the manuscript.

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.

*** We acknowledge that a t test would have sufficed, but in fact these two analyses are identical with the relationship that $t^2 = F$. Therefore, we don’t believe this distinction makes an important methodological difference although we have clarified the 2 independent-group nature of the test at line 2 of page 9 of the manuscript.

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?

*** To address this, it would be necessary to have a measure of stereotype activation which operates at an individual participant level rather than a group level. It is possible to calculate the mean difference between response times for stereotype words and neutral words for each participant, but this isn’t the way in which LDT data is usually handled, and whether it should be viewed as a continuous measure of stereotype activation is unclear (although admittedly theoretically logical).
If such a measure (referred to hereafter as “participant-level stereotype activation”) is calculated, then it is normally distributed, has a mean of 52.8ms, with sd of 44.5ms, and no statistically significant difference between experimental groups. When it is added as a co-variate to the analysis of either the focus or valence of the feedback, it shows no statistically significant relationship with the dependent variable, and its inclusion does not alter the significance of any of the main effects or interactions. In short, it has no important relationship to this finding.

We could present these data within the paper, but it does not relate to a theorised comparison and we know of no prior use of LDT data that would allow us to justify treating the difference in this manner. Additional explanation would be required to explain how the variable was calculated, and if it were to be described as a covariate for the feedback hypotheses, then it would seem logical to also present it in relation to the score and recollection hypotheses. This would therefore require several additions to the paper. It is our view that to do so won’t benefit the paper, however, if the editor judges it to be important then we would be happy to make this inclusion.

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

*** It is unclear to us exactly which aspects of this section were unclear. On reflection, the description of the expected theoretical pattern of results benefits from some elaboration, which has been added in lines 33 and 38 of page 10; if further clarification is required we would be happy to address this.

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 stated earlier, further consideration of these important theoretical discrepancies has been added in lines 4-12 of page 12, and additional consideration of the practical implications of the findings have been discussed in lines 13-16 of page 12.

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

*** This sentence has been broken up and clarified at lines 19-24 of page 11.

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.

*** Thank you. These typographic errors have been noted and amended.

Additional changes made by the authors:

After reflecting on our manuscript in response to the reviewers’ comments and with further consideration of our own we have made a number of additional minor amendments.

*** Whilst the development and validation of the stimulus materials and measures was described in the appendices (partly in response to reviewer 3’s comments regarding the content analysis of feedback) we decided to include very brief descriptions of salient aspects of stimuli development and validation in the main text, with more explicit references to the specific appendices which readers should consult for further details.

*** We decided that, on reflection, there was some room for ambiguity in our description of the way that erroneous responses were excluded from the LDT results. These have been clarified in lines 32-37 of page 8, with details of the proportion of responses which were actually excluded in lines 25-27 of page 9.

-------------------Editorial Policies-------------------

Please read the following information and revise your manuscript as necessary. If your manuscript does not adhere to our editorial requirements, this may cause a delay while this is addressed. Failure to adhere to our policies may result in rejection of your manuscript.

In accordance with BioMed Central editorial policies and formatting guidelines, all manuscript submissions to BMC Medicine must contain a Declarations section which includes the mandatory sub-sections listed below. Please refer to the journal's Submission Guidelines web page for information regarding the criteria for each sub-section (https://bmcmedicine.biomedcentral.com/).

Where a mandatory Declarations section is not relevant to your study design or article type, please write "Not applicable" in these sections.

For the 'Availability of data and materials' section, please provide information about where the data supporting your findings can be found. We encourage authors to deposit their datasets in publicly available repositories (where available and appropriate), or to be presented within the manuscript and/or additional supporting files. Please note that identifying/confidential patient data should not be shared. Authors who do not wish to share their data must confirm this under
this sub-heading and also provide their reasons. For further guidance on how to format this section, please refer to BioMed Central's editorial policies page (see links below).

Declarations

- Ethics approval and consent to participate
- Consent to publish
- Availability of data and materials
- Competing interests
- Funding
- Authors' Contributions
- Acknowledgements

Further information about our editorial policies can be found at the following links:

Ethical approval and consent:

http://www.biomedcentral.com/about/editorialpolicies#Ethics

Availability of data and materials section:

http://www.biomedcentral.com/submissions/editorial-policies#availability+of+data+and+materials