PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Order effects in high stakes undergraduate examinations: an analysis of five years of administrative data in one UK medical school
AUTHORS	Burt, Jenni; Abel, Gary; Barclay, Matthew; Evans, Robert; Benson, John; Gurnell, Mark

VERSION 1 - REVIEW

REVIEWER	Dr C M Wiskin University of Birmingham; UK
REVIEW RETURNED	18-May-2016

GENERAL COMMENTS	The paper is concisely presented, clear, and addresses a topic of interest across colleges. I enjoyed the paper (it actually re-confirms findings from research I undertook myself between 1997 and 2003), but please note the below. I have 2 concerns which need to be addressed please, and a series of minor modifications. While some reference is made to past work I think it is important to emphasise that this research question is not novel. I am aware of similar studies over the last 20 years, and there is I think quite a substantial (and widely accepted) evidence base confirming that OSCE repetitions over time or sites don't significantly disadvantage candidate groups. While not strictly a "limitation" the authors I think need to emphasise more that they are adding evidence to a known phenomenon. In line with that (my second point) the literate reads as rather dated. I think some more recent references would help the content. I am certain that psychometric work exists after 1989…., and there are fairly robust UK examples (add to P5). Regarding the text, for your consideration: Pm 3 line 12 refs 1-2 should come after OSCE, not after 'sills' P 3 line 17 'OSCE examinations2 is just 'OSCE' (the E being examination…) P 3 after line 32, add perhaps that question rotation is also a means of helping test security. P 3 line 48 'a candidate's performance' (not 'a candidates') P4 line 22 OSCE abbreviation has already been explained in the Intro? P4 line 58, is the 50% of stations needing to be passed correct? Just checking as seems unusually generous for finals P5 line 18-19, if a pass rate on only half the stations is all that is expected the 98/5%passrate is not surprising (given the low threshold for success) – is it work mentioning this? P 5 line 41 I think loose "highly" (its either significant, or not, so 'shades of significance' are not usually reported?) P7 line 13 is ther4e any concern over the very high pass rates emerging from this data? Does the pass rate merit revisiting

	standard setting at all? Examiners are referred to as potential confounding factors. Did you also consider SPs, patients? Tables seemed clear/relevant. Overall a nice paper. Thank you.

REVIEWER	Amir Sam Imperial College London
REVIEW RETURNED	20-Jun-2016

GENERAL COMMENTS	This is a well written manuscript, addressing an important and pertinent question, particularly in view of the discussions about introduction of the UK Medical Licensing Assessment. Some points for authors to consider: 1. It would be nice to show that there was no significant correlation between the order in which the students sat the exams and students' rankings based on each exam. 2. To what extent did the exam content change between days? Is it possible to look at the performance in the stations that were repeated between days and those that were varied/different? 3. How does the Paediatric question bank compare to that of the SCEE? Is it possible that there are less questions in this exam compared to the SCEE? Does the ratio of simulated/real patients differ between the Paediatric OSCE and SCEE? 4. How did the cohorts in various years compare in the written SBA assessments?

VERSION 1 – AUTHOR RESPONSE

Reviewer 1: Dr C M Wiskin, University of Birmingham

“While some reference is made to past work I think it is important to emphasise that this research question is not novel. I am aware of similar studies over the last 20 years, and there is I think quite a substantial (and widely accepted) evidence base confirming that OSCE repetitions over time or sites don’t significantly disadvantage candidate groups. While not strictly a “limitation” the authors I think need to emphasise more that they are adding evidence to a known phenomenon. In line with that (my second point) the literature reads as rather dated. I think some more recent references would help the content. I am certain that psychometric work exists after 1989…., and there are fairly robust UK examples.”

In response to the reviewer’s comments, we repeated our search of the literature to locate further relevant studies. As a result, we identified two additional papers we had not cited; a study from 1991 looking at potential violations in test security in standardized-patient cases at one US medical school (Colliver, Barrows, Vu et al. 1991) and a 2003 study looking at discussions of OSCE exams by students on an electronic discussion board (Parks, Warren, Boyd et al., 2003). We have now added references to these within the introduction to the paper. However, we were unable to locate any more recent relevant literature, and would welcome any further suggestions from the reviewer in this area.

Minor modifications:

P3 line 12 refs 1-2 should come after OSCE, not after ‘sills’
We have amended accordingly

P 3 line 17 ‘OSCE examinations’ is just ‘OSCE’ (the E being examination…)
We note this and have amended accordingly (there was more than one instance of this, which we have now corrected)

P 3 after line 32, add perhaps that question rotation is also a means of helping test security.
We thank the reviewer for highlighting this issue. On reflection, we have decided against adding further detail to the manuscript, as our analysis is focused on OSCEs where questions are identical in sequential candidate groups. In response to this, and a comment made by reviewer two, we have, however, clarified the situation within the manuscript by adding the following text to the methods section:

“For all OSCEs, the content of each station, including the question wording, did not vary between circuits or between days.”

P 3 line 48 ‘a candidate’s performance’ (not ‘a candidates’)
Corrected

P4 line 22 OSCE abbreviation has already been explained in the Intro?
Thank you – we have deleted this explanation

P4 line 58, is the 50% of stations needing to be passed correct? Just checking as seems unusually generous for finals

We have clarified this section of the paper to stress that candidates must meet two criteria to pass the exam: (a) meet the overall examination pass mark and (b) pass 50% of stations – the amended text now reads:

“In the case of all three examinations, in order to pass, students were required (a) to meet the overall examination pass mark, as defined by the borderline group method, 13 and (b) to additionally pass a minimum of 50% of individual stations: this ensures that poor performance in several stations cannot be compensated for by exceptionally high performance in one or two other stations.”

We note that these conditions are approved by the University of Cambridge: a high pass rate in finals examinations is to be expected, and is in line with that seen in other institutions. The current analysis is not commenting on the overall pass rate, but comparing pass rates in sequential cohorts.

P5 line 18-19, if a pass rate on only half the stations is all that is expected the 98/5%passrate is not surprising (given the low threshold for success) – is it worth mentioning this?

See our response above – candidates are required both to pass 50% of stations and to meet the overall examination pass mark, and we have clarified this in the text.

P 5 line 41 I think loose “highly” (its either significant, or not, so ‘shades of significance’ are not usually reported?)

After careful consideration, we must respectfully disagree with the reviewer on this point – i.e. that results are either significant or not. Such an approach ignores the continuum of evidence that p-values provide. Thus, p=0.001 provides much stronger evidence against the null hypothesis than p=0.05, and we believe this should be made clear in the text. Whilst we are aware that ‘shades of significance’ are not always reported, we believe there is justifiable merit in this practice.

P7 line 13 is there any concern over the very high pass rates emerging from this data? Does the pass

rate merit revisiting standard setting at all?

The Cambridge Final MB Exams are subject to independent review by a panel of External Examiners, who cover all disciplines with respect to their expertise. They have confirmed year on year that the Cambridge Exams are a rigorous assessment of knowledge, skills and competence. We note that Cambridge is fortunate to benefit from very high achieving students – this is also supported by evidence from the postgraduate arena which shows that Cambridge students are amongst the highest achievers in postgraduate examinations. Thus, we do not have concerns over the pass rates shown in our data.

Examiners are referred to as potential confounding factors. Did you also consider SPs, patients?

We acknowledge that the SPs or patients for each station vary between circuits and days, and thus may be a confounding factor. However, it was not possible to investigate the impact of SPs or patients as we do not hold data on the participants for each station other than the candidate and the examiner. We note, however, that all SPs are trained to a high standard, and the lead for SPs together with the SPs participating in any given station discuss the station in detail in advance and agree how they will respond to particular student questions or approaches – thus Cambridge actively seeks to minimise variation in SP performance, although we acknowledge that this cannot be eliminated completely. We have added the following text to the discussion section of the manuscript to address this point:

“Additionally, the simulated patients and patients used within the examinations may vary between circuits and days; we were not able to investigate the potential impact of this, as we did not hold information about the simulated patients and patients involved in these examinations. However, we note that all simulated patients are trained to a high standard and discuss each station in detail in advance in order to minimise variation in performance.”

Reviewer 2: Amir Sam, Imperial College London

1. It would be nice to show that there was no significant correlation between the order in which the students sat the exams and students' rankings based on each exam.

This paper considers the impact of examination grouping to which students were assigned on examination performance. In our analysis, we were not investigating whether overall order made a difference; however, if it did, we would expect it to have been highlighted in our analysis of the day that the exam was started (models 1 and 2). To clarify this, we have added the following text to the discussion section:

“We did not seek to see if overall order of examinations taken made a difference; however, if this had been the case, we would expect it to have been highlighted in our analysis by the day the exam was commenced.”

2. To what extent did the exam content change between days? Is it possible to look at the performance in the stations that were repeated between days and those that were varied/different?

We can confirm that the exam content did not change between days. Each station was repeated on each day of each exam, and for any given station approximately the same number of candidates were assessed on that station on each day of each exam. This means we could not investigate the impact of exam content changes between days. In response to this comment and a related comment made by reviewer one the following text has been added to the methods section:

“For all OSCEs, the content of each station, including the question wording, did not vary between

circuits or between days."

3. How does the Paediatric question bank compare to that of the SCEE? Is it possible that there are less questions in this exam compared to the SCEE? Does the ratio of simulated/real patients differ between the Paediatric OSCE and SCEE?

The paediatric and O&G station banks are slightly smaller than that of the SCEE; and each paediatric or O&G examination has 8 or 9 stations compared to the 10 of the SCEE. We do not believe these factors impact on the current analysis.

With respect to the ratio of simulated/real patients, there are a maximum of two or three real patients in the paediatrics exam, and one in the O&G exam: in response to reviewer one we have added the following statement to the discussion to acknowledge any potential variation which may be attributed to patient factors:

"Additionally, the simulated patients and patients used within the examinations may vary between circuits and days; we were not able to investigate the potential impact of this, as we did not hold information about the simulated patients and patients involved in these examinations. However, we note that all simulated patients are trained to a high standard and discuss each station in detail in advance in order to minimise variation in performance."

4. How did the cohorts in various years compare in the written SBA assessments?

We have not formally performed such an analysis as we believe that, although potentially interesting, it is not directly relevant to the question asked in our study, which was to investigate the association between the scores achieved by students and the examination grouping to which they were assigned. We are concerned that the addition of such an analysis might distract/detract from the key message of the paper.

VERSION 2 – REVIEW

REVIEWER	Dr C M Wiskin College of Medical & Dental Sciences University of Birmingham UK
REVIEW RETURNED	04-Aug-2016

GENERAL COMMENTS	As this is a re-submission with modest revisions I will not review at length. My response here is to the authors comments, which in turn replied to my earlier and more substantive review. The questions I raised have, I feel, been addressed and I am happy to see the current draft published. In the one or two instances where we have different views, I accept the authors' desire to retain their own working, and their rationales. Thank you for the opportunity to see the revised paper, and good luck with any ongoing or new endeavours.

REVIEWER	Amir H. Sam Imperial College London, UK
REVIEW RETURNED	18-Aug-2016

GENERAL COMMENTS	I am happy with the authors' reply.