Reviewer’s report

Title: Investigation of Answer Changes on the USMLE® Step 2 Clinical Knowledge Examination

Version: 0 Date: 10 Jul 2019

Reviewer: William Howard Adams

Reviewer's report:

The authors of an "Investigation of Answer Changes on the USMLE Step 2 Clinical Knowledge Examination" have impugned (in their words) the "conventional wisdom" that medical students should avoid changing their answers on the USMLE Step 2 Clinical Knowledge examination. In their timely investigation, they find that not only do most examinees change their answers, these changes are often from an incorrect response to a correct response. Further, they show that such changes vary by examinee ability and item difficulty. This manuscript is generally well written and clear, and I have only a few questions regarding the statistical analysis plan for their consideration.

Minor Essential Revisions:

1. Methods Statistical Analyses: For any given item, the authors have excluded examinees who do not change their response. That is, they only retain examinees who have changed an item from right-to-wrong-to-right (RR), right-to-wrong (RW), wrong-to-right (WR), and wrong-to-wrong (WW). Examinees who marked a correct response and never deviated from it (R) were excluded (from that item) as were examinees who marked an incorrect response and never deviated from it (W). The reasons for excluding these response patterns is not clear to me, particularly in the authors' regression models where they regress examinee ability (i.e., high, medium, and low) and revision type (i.e., WR, RW, RR, and WW) on the dependent variables of elapsed item time (Table 5) as well as item difficulty (Table 6). Why not spend the extra degrees of freedom and include these response patterns in the models? At the very least, you would be using all of the data which has an impressive sample size. Or, alternatively provide a compelling reason for dropping them from the analysis.

2. Methods Statistical Analyses: It seems to me that the independence assumption of the analysis of variance (ANOVA) model may be violated. For example, in Table 5 the authors show mean item duration (in seconds) in a four-by-four contingency table where the rows are defined by the examinees' variable response pattern (i.e., WR, RW, RR, and WW) and the columns are defined by the total sample as well as three subgroups of examinee ability (i.e., high medium, and low). From my reading of this Table as well as the results section, it appears that each examinee is contributing multiple items to the analysis. If that is true, then the independence assumption of
the ANOVA model is clearly violated. This is because each examinee is contributing multiple items to the analysis and, therefore, he/she is contributing multiple item time stamps to the analysis. For this reason, the authors may wish to consider using a linear-mixed effects model rather than an ANOVA model. Such a model could include a random intercept for each examinee to account for the correlation of elapsed item times within an examinee. This seems reasonable, as an examinee who spends a considerable amount of time changing their response on one item is likely to spend more time changing their responses on other items. This correlation, if present in the outcome, needs to be appropriately modeled. This also applies to Table 6.

Discretionary Revisions:

3. Methods Statistical Analyses: I ask the authors to consider swapping the outcomes in their regression models. That is, the authors currently evaluate whether the dependent variables elapsed time (Table 5) and item difficulty (Table 6) vary by students' variable response pattern (i.e., WR, RW, RR, and WW) and examinee ability (and presumably their interaction). But that seems strange to me. Would it be more useful to measure students' variable response pattern as the outcome (rather than elapsed time and item difficulty)? To me, the variable response pattern is the star player in this otherwise excellent report and should take center stage as the outcome (dependent) variable of interest (not elapsed item time or item difficulty). The authors could consider regressing elapsed item time and student ability on the probability of a WR, RW, and/or RR response (against the referent of a WW response). With such a model, the authors would be able to say something along the lines of "Controlling for elapsed item time, the odds of a high ability examinee marking a WR (rather than WW) response are XX (95% CI: XX - XX) times higher than a low ability examinee. Similarly, compared to low ability examinees, high ability examinees were XX (95% CI: XX - XX) times more likely to mark a RW (rather than WW) response, and they were XX (95% CI: XX - XX) times more likely to mark a RR (rather than WW) response." And so on. The authors could accomplish this analysis using a generalized linear mixed-effects model that specifies a multinomial distribution with generalized logit link for the four-level nominal outcome (i.e., comprising the response patterns WR, RW, RR, with referent WW). This also applies to Table 6.

Minor Issues not for Publication:

4. The results for Block 8 should be included in this good work. The fatigue effect, if any, should be shown. Even if these results are submitted as supplemental content, we should see them.
5. The scale for the IRT item difficulty parameters in Table 6 should be provided. Usually IRT difficulty estimates range from negative infinity to positive infinity and are scaled scores analogous to z-scores (with a mean of 0 and SD = 1). What is the scale here?

This was a well-written study evaluating student ability and their variable response patterns on the dependent variables of elapsed item time (Table 5) and item difficulty (Table 6). I think this work would be better if the outcome was the variable response pattern, which is the major outcome of this work. Further, if the data structure is ultimately hierarchical such that examinees contribute multiple items to the analysis, then a mixed-effects model or marginal model is needed to account for the hierarchical sampling structure.

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

No

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

No

**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/). 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.