Author’s response to reviews

Title: Interprofessional assessment of medical students' competences with an instrument suitable for physicians and nurses

Authors:
Sarah Prediger (s.prediger@uke.de)
Sophie Fürstenberg (s.fuerstenberg@uke.de)
Pascal Berberat (berberat@tum.de)
Martina Kadmon (martina.kadmon@med.uni-augsburg.de)
Sigrid Harendza (harendza@uke.de)

Version: 1 Date: 19 Dec 2018

Author’s response to reviews:

Dear Dr. Messin,

Below please find our point-by-point reply to all reviewers’ comments and suggestions. Thank you, they were very helpful to improve our manuscript. Changes in the manuscript are marked in yellow. We hope that with these revisions our manuscript might now be acceptable for publication in BMC Medical Education.

With kind regards,

Sigrid Harendza

Reviewer reports:

Jocelyn Lockyer (Reviewer 1):

BMC. Interprofessional assessment of medical student competencies.
This is an interesting study. It could be helpful to others attempting to assess interprofessional competency. There are a few aspects that are not entirely clear and would benefit from some editing.

Abstract.

In methods, might you identify the 7 competencies?

Reply: All seven facets of competence are identified in the abstract now.

In results, you have provided information about the nurse correlation of competence with confidence but not for residents. For supervisors, might you identify which two competencies were correlated with confidence?

Reply: With the inclusion of the seven facets of competence in the abstract, as stated above, we have reached the word limit of 350 words for the abstract. Therefore, this aspect cannot be included in the abstract.

In conclusions, it seems that the final sentence (Rating confidence) goes beyond the study findings. I wonder if it might be more appropriate to summarize the key findings related to the two aims of the study in Conclusions.

Page 2, line 61 and 64 should be competencies, not competences.

Reply: Rating confidence was one of our two aims of the study and we were able to improve rating confidence at t2 by including anchors in the rating forms. Therefore, we would like to keep the final sentence, which refers to this aim of our study. Since we use the words competence and competences throughout the manuscript (and not competency and competencies) we would like to stick to competences on page 2, line 61 and 64.

Background.

The literature review is helpful and comprehensive.
The aim as expressed at the end of the introduction seems inconsistent with the description in the abstract and what was done in the actual study. Would it be more accurate to say that the aim of the study was to (1) compare the ratings provided on different competencies by three different assessor groups and (2) assess the association between the scores on the assessed competencies and the confidence in judgement by the three assessor groups?

Reply: We aligned the two aims of our study at the end of the background section closer to the aims as they are expressed in the abstract. The first aim of our study was to compare the three different assessor groups with respect to their rating scores for the different facets of competence to identify, whether the assessment instrument is suitable for the three rater groups. Our second aim was to analyse the association of assessment scores in comparison with the confidence of rating within each rater group, to provide further insight into rater cognition.

Methods.

Did the actual instrument consist of 7 items? Or were there multiple items used to explore each of the competencies? Could the actual instrument be provided? It appears there was a lot of discussion (ie, training on how to use the instrument) and it would be useful to have the salient details of what was to be assessed included (beyond the 7 competencies as listed on pages 6 and 7 (lines 154 - 157).

Reply: Your impression is correct; the FOC-scoring form consists of seven items, which refer to the seven facets of competence. Each facet of competence includes a definition that could be used as guideline by the assessors for observing performance. We included a sample of the assessment form in appendix 1.

Can you clarify what you mean by 'discuss their judgements of situations in movies to build shared mental models? (page 7, lines 165-166).

Reply: All raters were trained a few weeks before the assessment to learn how to use the rating instrument. They rated two videos of an excellent and a mediocre student taking a simulated patient’s history to build shared mental models for each facet of competence. According to Oudkerk Pool et al. (2017), mental models affect judgement decisions. Therefore, our training was used for the standardization of the raters.
What do you mean by semester 10? (page 7, line 170)

Reply: We replaced semester 10 by year five of undergraduate studies.

Can you explain what the effect size calculation was intended to do and why it is necessary in addition to the p-values associated with the ANOVA and Bonferroni post-hoc test.

Reply: While the p-values only explain the significance of the differences, i.e. how high the percentage is that the finding is incorrect (usually p<.05, which means that there is less than 5% possibility that the finding of a difference is incorrect), the effect size is a measure how large the significant difference between groups actually is and whether it is relevant. Hence, it is important to show in our study, whether the significant difference have actual practical relevance or not.

Results.

For the internal consistency (α), the results should be provided without a comment on whether it was satisfactory or not. Are you not providing the internal consistency of the instrument by group?

Reply: Yes, the internal consistency of the instrument is provided for each rating group. The indication of “satisfactory” was deleted from the sentences.

Confidence of judgement is presented as a figure. Given the types of analyses you described, would it not make more sense to provide the confidence information in a table format. It might be included in Table 1 or as a separate table. You can then proceed to table 2 (as described on page 9, lines 210-215).

Reply: We now provide the means of the confidence of judgement in a new table instead of presenting them in figure 2.

Discussion.
In the discussion, I would recommend that you go back to the aims for the study. Do some editing related to the key findings for each aim and then discuss the aims in conjunction with the broader literature.

Reply: According to the reviewer’s suggestion we divided the discussion in two sections which each starts with the key findings for aim one and aim two, respectively. Each finding is then discussed in conjunction with the broader literature.

For limitations, consider whether your sample sizes were large enough to do the calculations and draw the conclusions that you did? I wondered whether the limitations as described were actually implications for future studies.

Reply: The sample size of our assessment was 70 participants. According to the limit theorem, a Gaussian distribution can be expected when the number of participants exceeds 30. Hence, we regarded our statistical analyses as feasible. We added a comment in the limitations section. Despite of the low sample size, we were able to identify significant differences between rating groups and significant correlations between assessment scoring and confidence of judgment as well as high effect sizes. The limitations we found were used for conclusions how to further improve the assessment. Even though limitations cannot always be avoided they may be helpful for improvement. We comment on this in the limitations section.

Wolf Hautz (Reviewer 2): Dear authors,

thank you for the opportunity to review your manuscript on Interprofessional assessment of medical students' competences.

I applaud you for conducting such an extensive assessment with students and assessors from three medical schools, which I believe to pose a rather substantial logistical challenge. What adds to the relevance of the paper is the fact that it is still very much unknown how competences should be assessed, especially those you refer to as "specific non-medical".

To further improve your paper, I have several suggestions.
Introduction:

Line 104ff: Please be more specific here. I would suggest to name a few of the conditions which need to be met to achieve satisfactory psychometric characteristics (including reliability) in an OSCE. I think the claim that "the students are assessed at different stations by more than one rater, mostly physicians, [...] leads to good reliability" per se is too broad. The AMEE guide you cite gives a number of conditions which should be met to achieve satisfactory reliability in an OSCE.

Reply: We added additional criteria, which lead to more reliable and valid scoring outcomes of OSCE by citing an important meta-analysis (reference 13).

Lines 110ff: I would suggest differentiating between assessment purposes. While OSCEs are often used in summative exams, where reliability is important, I think there is a general agreement in the literature that work place based assessments should only be used for formative purposes (Frank Med Teach 2010, Schuwirth Med Teach 2011, Lörwald Med Teach 2017), precisely because they are rather unreliable.

Reply: We included a sentence to point out the different assessment purposes and cited two of the suggested references.

Research question, line 133ff: Your introduction leads to the first aim of your study, namely the question if different assessor groups assess candidates differently. The second part (how confident assessors are in their ratings) however comes as a bit of a surprise. I think it's a quite innovative aspect to look at, but I would suggest to provide a little bit of a rational as to why you are interested in assessor confidence. I am not aware of any research in that field, but the literature on confidence judgements in learning and clinical reasoning (see eg. Cue utilization framework by Anique de Bruin, Med Educ 2017, or a recent review by Ilgen et al in ASHE 2018, the work of Asher Korriat or Eva & Regehr on self Monitoring) cold provide starting points. A rational as to why you assess confidence, together with a hypothesis, could also help to frame the discussion of this particular aspect.

One could for example hypothesize that when confident, raters tend towards bolder judgments (more towards the borders of the scale) whereas when unconfident (but forced to provide a score anyway) they might be more lenient (i.e. scoring towards the scales mean).
Reply: To assess raters’ confidence of judgement was already part of our initial pilot study of this simulated assessment with Olle ten Cate (Wijnen-Meijer et al. Adv Health Sci Educ Theory Pract 2013). For the instrument to work, i.e. to assess the facets of competence, they have to be visible for the assessors by the candidates’ performance during the simulation. Since, as you are pointing out, our assessors were also given the choice of marking “assessment not possible” when they felt they were not able to observe a certain facet of competence – and they were trained to do so – we received feedback what to change in the simulation conditions to make a certain facet of competence observable. Furthermore, we achieve a better understanding of aspects influencing assessment outcome of our new scoring instrument. Analyzing the association of scores for a candidate with the confidence of the ratings discovered an important additional aspect in rater cognition. One of the conclusions from the confidence of rating results was, that the definitions of the facets of confidence are not sufficient for confident ratings. Hence, we suggest to provide anchors for each facet of competence. We made this more prominent in the introduction and highlighted it in the conclusions.

Methods:

Lines 153ff: I would suggest to explain how you came up with the 7 competencies you assess, in particular because in the introduction (line 97), you claim that "different competences should be assessed separately, to give raters the opportunity to focus on each competence individually".

Reply: The seven competences were extracted from the top 10 facets of competence from a ranking study of 25 facets of competence according to their necessity for a beginning resident (please refer to the first sentence of the methods section and its reference. We clarify now in the introduction, what we mean by “assessing separately”. Competence is not assessed as one construct, but rather certain facets of competence, which can be assessed by observing performance, are assessed separately (like the seven facets of competence used in our study). To avoid confusion, we added “facet of competence” in the whole manuscript when necessary.

I would further suggest extending the analyses. It is interesting that nurses score more lenient than the other assessors. One important question however is whether this simply reflects a shift of scales, or whether the rank order of students changes depending on the assessor. In other words: can we convert the supervisors score into a nurses score simply by multiplying it with a fixed factor and those that score high in the supervisors assessment also score high with the nurses, or are the best/worst performers actually different depending on assessor.
Reply: Thank you very much for your suggestion to look at this additional analysis. We adjusted nurses scoring by reducing it according to the delta between supervisors’ mean score and nurses’ mean score for an individual participant. The comparison of adjusted nurses’ scores and supervisors’ scores in an individual ranking per facet of competence showed an average agreement of 42.8 % of both rating groups, with higher agreements for participants, who were assessed “good” or “very good”. We added this information in method and results and also discuss it for further developments. This finding supports the finding for OSCEs, that interpersonal skills were evaluated less reliably across stations (Brannick et al. Med Educ 2011). However, the same facet of competence, e.g. teamwork, could be differently performed by the same participant with respect to interacting with a supervisor or a nurse and lead to different scores because of the different perspectives. As we already concluded, it will be important to develop rating anchors for the different facets of competence, which can also be used to give feedback to the participants and which will help to explain, why a facet of competence might be rated differently by supervisors and nurses for the same participant.

Results:

Line 191: Please substitute "p<0.05" with "all p<0.05".

Reply: We have substituted "p<0.05" with "all p<0.05".

Line 189: I was surprised to see that you calculated alpha for your assessment. If I understand correctly, you propose that your assessment measures 7 different competencies. Then why would you calculate only one alpha? In theory, the 7 competencies you assess should be (at least) partially be independent of each other (because otherwise, they would be just one competence). A factor analysis could tell.

Reply: The correlation analysis of our seven facets of competence shows significant correlations between most of the facets. The factor analysis, which we had already conducted, showed two factors with some facets of competence, which could not clearly be assigned to one or the other factor, because they loaded equally on both factors. Therefore, we can assume, that these facets are interrelated and this might also be reflected in the rating scores. We also calculated alpha for each facet of competence for particular rating of patients/disturbances with the total rating. This showed equal alpha-scores compared to reported alpha scores per group in the manuscript.
Therefore, we decided to report only the alpha scores for each group over all facets of competence in the manuscript. Since the facets of competence comprise very different aspects (e.g. “teamwork and collegiality” and “structure, work planning and priorities”) it is important for the individual participant to receive feedback about his or her individual performance with respect to their scores for the different facets of competence to know, which facet needs improvement (e.g. somebody might work well in a team but might not be very good at structuring his or her work and at setting priorities).

Discussion:

As far as I understand, assessors are obliged to provide a score (unless they did not observe the respective behavior at all). It would be interesting (from my point of view), whether it is their confidence in their assessment that predicts their score or whether their score informs their confidence. Please see my comment on confidence above: I would suggest to extend the discussion of the relationship between score and confidence in it to the cues that may inform either.

Reply: This is a very interesting question, whether confidence predicts the assessment score or vice versa. Unfortunately this question cannot be answered by our correlation analysis. As we argue in the discussion, we interpret our findings that confidence is related to the observability of competence and that anchors are needed to increase rating confidence. Whether this is the case will be tested in the next round of our assessment.