Author's response to reviews

Title: Evaluation of Large-group Lectures in Medicine - Development of the SETMED-L (Student Evaluation of Teaching in MEDical Lectures) Questionnaire

Authors:

Tjark Mueller (tjark.mueller@gmx.de)

Diego Montano (diego.montano@med.uni-goettingen.de)

Herbert Poinstingl (herbert.poinstingl@med.uni-goettingen.de)

Katharina Dreiling (katharina.dreiling@med.uni-goettingen.de)

Sarah Schiekirka-Schwake (sarah.schiekirka@med.uni-goettingen.de)

Sven Anders (s.anders@uke.de)

Tobias Raupach (raupach@med.uni-goettingen.de)

Nicole von Steinbüchel (nvsteinbuechel@med.uni-goettingen.de)

Version: 1 Date: 27 Apr 2017

Author’s response to reviews:

Dear Prof. Azer,

Thank you for inviting us to submit a revised version of our manuscript. We found the editor’s and reviewers’ comments very useful and believe that they have helped to further improve the manuscript. We used track changes to highlight new additions to and deletions from the manuscript text. Please find a point-by-point reply to all comments below.

We hope you will now find our manuscript amenable for publication in BMC Medical Education.

Kind regards,

Tobias Raupach & Sven Anders on behalf of all authors

Editor Comments
1) The paper cannot be published in its current format. If the authors are interested in submitting an amended version, they should address all points raised by each reviewer and make these changes in a different color (Red or Blue) and underline these changes. They should also submit a letter to the Editor addressing in a point-by-point format each point raised by each reviewer, state their response, and indicate page number and lines, in the manuscript, of these changes.

2) Reply: We have tried to comply with all of these requirements.

Reviewer 1

1) This is an interesting paper that attempts to validate a brief questionnaire that can be applied towards the assessment of large group lectures in a medical school. As a primary educator, I found the paper to be very heavy on the statistical methods and would have appreciated, early in the paper, perhaps a concise and convincing statement of "what we will miss" if we don't have a validated and user-friendly assessment tool such as the one that they attempted to test.

Reply: We agree that statistics are featured quite heavily in this paper. However, given that methods have moved on considerably in the past few years we felt it was necessary to include some detail for readers who are not too familiar with our analytical approaches as psychometric quality of instruments is crucial for reliable and valid research and routine use. We would like to support readers in making informed choices about which methods they will use in their own future work. At the same time, we believe that this reviewer comment is important because not all readers will be interested in these details. In order to put more emphasis on practical implications, we have added the following paragraph to the introduction:

“Most available questionnaires focussing on teaching quality and teacher characteristics are quite comprehensive, thus precluding their routine use. However, specific feedback on teaching quality for all lecturers involved in a course is needed for two reasons: First, improvements of teaching quality are much easier to achieve if lecturers can use reliable and truthful data to identify specific aspects of their teaching that can be improved. Second, as medical schools make increasing use of evaluation data to inform decisions on individual careers, reliable and valid evaluations are required for individual teachers in order to render these decisions as fair as possible.”

2) I feel that it would have also been helpful in the discussion to state, in practical terms, how such a form can be implemented in a medical school curriculum and the expected benefits from using this tool (that could not otherwise be achieved).

Reply: Thank you for this helpful suggestion. We have included a new section labelled ‘Practical implications for evaluation in medical education’ into the discussion:

“To our knowledge, the SETMED-L is currently the shortest questionnaire evaluating teacher performance in medical lectures. In addition to its items being derived from a widely-used
framework of high-quality teaching, its psychometric properties are favourable, and the analyses regarding measurement invariance suggest that student gender does not affect evaluation results. In addition, few instruments apart from the SETMED-L include items of student activation. Due to these advantages over existing questionnaires, the SETMED-L lends itself to routine use in courses with multiple lecturers. Curriculum-wide implementation of the tool will provide a rich database from which teacher rankings may be derived and these evaluations may result in informed decisions on individual career pathways.

Apart from this summative function, one major purpose of the SETMED-L is to provide formative feedback to teachers thriving to further improve their didactic skills. Information on specific aspects of teaching can be used to tackle corresponding areas of potential improvement. From a faculty development perspective, lecturers who have received their individual evaluation results could be invited to participate in teacher trainings tailored to their specific needs, and repeated measurements following training could be used to assess the progress made. Qualitative studies assessing teacher and student perceptions of the new tool following curriculum-wide implementation are currently under way.”

Reviewer 2

1) This is an interesting paper describing the development of an evaluation tool for lectures in medicine. The importance of feedback to teachers for the enhancement of students' learning has received increasing attention. Standardized and validated instruments are preferable to locally developed questionnaires and this paper presents the development of such an instrument. The study is well designed and performed and I only have a few comments.

Reply: Thank you for this positive appraisal of our work.

2) Page 5. I wonder why students were not involved in the development of the instrument items. Since the students are the ones who will answer the questionnaires it is important how they interpret the items and how they judge their relevance. I believe this is a weakness of the study that should be discussed in the paper.

Reply: Upon reexamining the paper we felt that parts of the methods section might have been confusing as the analytical approach was partially reported in sections primarily referring to item generation or data collection. This might also be the reason why student involvement was not apparent from the manuscript in its prior form. We have now rearranged the methods section and explain in more detail how students were involved at the stage of item generation and wording (cognitive debriefing) during the pilot phase. To this end, we have included a new paragraph entitled ‘Pilot testing and student involvement’.

3) On page 4 a pilot test with 898 students is mentioned and on page 7 246 students completing the FEVOR. However, the collection of data for studies 1 and 2 are not mentioned in the Methods section. This is somewhat confusing and I suggest that they should be included.
Reply: We totally agree and are sorry for omitting this important issue. This has now been corrected. The corresponding parts of the manuscript read:

Pilot testing: ‘In a consecutive pilot testing in the 2014 summer term, the 27-item version was handed out to 898 students at the medical schools of Göttingen and Hamburg.’

Study 1: ‘In winter term 2014/15, a total of 922 students evaluated teachers delivering large-group lectures at Göttingen or Hamburg Medical School (Study 1).’

Study 2: ‘Data collection for Study 2 took place in summer term of 2015 and winter term 2015/2016 at both medical schools. A total of 2740 student ratings were obtained.’

4) A description of how the instruments were distributed should also be included - paper at the end of lectures, electronic survey etc?

Reply: This information is now given at the end of the section on data collection:

‘Paper versions of the questionnaires were distributed to students and completed during the final 5 minutes of the lecture. All participating lecturers agreed to allocate this time for questionnaire completion.’

5) Page 6, line 27 states that the questionnaire version was subjected to cognitive debriefing, but it is not described how. An addition of this information is needed.

Reply: This comment relates to comment 2 (see above). The following description has been included in the methods section:

‘In cognitive debriefing sessions students were invited to discuss questionnaire instructions, items and answer scales with regard to content and wording in a formalized detailed way to avoid potential misinterpretations. Also, potentially missing questions were discussed; however, none were identified. A total of 30 students participated in this process, and some item wordings were revised according to their comments.’

6) In the results section pages 10-12 the results concerning how many students completed the questionnaires are described. But it is not mentioned how they may reflect the total student population. I understand that this is difficult but an indication of the response rate and whether e.g the gender distribution matched the total population could be included.

Reply: This is in fact difficult, but comparisons with total enrollment figures and an earlier study suggest that our samples are unlikely to be severely biased. However, given that selection bias is a potential threat to validity of our study results, this limitation is also mentioned in the discussion as follows:

‘In Study 1, information on gender was missing in 5.7% of cases. The observed distribution (41.8 % male; 52.5 % female students) largely reflects current enrolment data at both medical
schools involved in this study. A recent national survey including almost 20,000 German medical students yielded rates of 35% (male) vs. 65% (female) [Strobel et al. Addiction 2012].

7) Page 14, line 10, would not lecturer be better word than tutor here? A tutor is usually associated with small group learning.

Reply: We agree. In order to reduce redundancies (the word ‘lecture’ appears shortly after ‘tutor’ in this sentence), we changed ‘tutor’ to ‘teacher’.

8) Page 15 line 24 etc, the item about appropriate workload is interesting when the lecturer wants to use the results for development of the lecture. It the grading of this item is low, is the workload too high or too low? This could be discussed.

Reply: The corresponding section reads:

‘While on the FEVOR questionnaire, students are required to quantify the workload between "too little" and "too much", the items of the SETMED-L asked students to rate the appropriateness of the workload between "strongly disagree" and "strongly agree", and thus they were not discriminating the directions of inappropriateness explicitly.’

Our decision against using a low/high grading was due to the fact that we wanted to use the same scale for all items. As a consequence, lecturers cannot discriminate the direction of inappropriateness. This may be a shortcoming, but at the same time most teachers will be able to critically assess whether the workload was too high or too low. If they are completely lost, one option would be to ask a few students whether they thought too much or too little content was covered in a lecture.’

9) Finally I would also like to see a discussion on the future use of the instrument. If it has been developed for formative use and not mainly summative, the consequences of its use are a very important aspect of the validity. How do the lecturers react to and interpret the results? Do they know what to do if the scores are not entirely positive? See e. g. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 2006 Feb;119(2):166.e7-16.

Reply: Please see our reply to comment 2 provided by reviewer 1. Given that both reviewers suggested to include more information on practical implications (including the distinction between summative and formative use), we have written an entirely new section. At the end of that section, we refer to studies assessing the tool’s acceptability to students and teachers. These studies are currently under way.

10) There are some minor issues with the English, especially in the introduction (lines 19, 54, 59).

Reply: We are sorry for these language errors and have attempted to improve the quality of our written English. To this end, we asked a native speaker to check the manuscript, and she
suggested a few changes to the item wordings reported in Table 1. Finally, we have revised the title of the questionnaire (SETMED-L instead of EMT-L).