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)	Protocol for a randomized trial evaluating the effect of applying gamification to simulation-based endoscopy training
AUTHORS	Scaffidi, Michael; Khan, Rishad; Walsh, Catharine; Pearl, Matthew; Winger, Kathleen; Kalaichandran, Ruben; Lin, Peter; Grover, Samir

VERSION 1 – REVIEW

REVIEWER	Andreas Slot Vilmann CAMES-Rigshospitalet Denmark
REVIEW RETURNED	28-May-2018

GENERAL COMMENTS	This is a very well written and comprehensive protocol. I have only minor comments. Firstly, when did the inclusion of participants start? and secondly how will the authors approach a potential sex related difference in performance? Former studies in simulation based training indicates a performance gap between women and men favoring the latter (Surg Endosc. 2015 Nov;29(11):3065-73. doi: 10.1007/s00464-015-4092-2. Epub 2015 Jan 29. "Gender differences in the acquisition of surgical skills: a systematic review." Ali A1, Subhi Y2,3, Ringsted C4,5, Konge L2.)

REVIEWER	Simon Blackburn Great Ormond Street Hospital for Children, UK
REVIEW RETURNED	25-Jun-2018

GENERAL COMMENTS	Thanks for asking me to review this manuscript. This is a well described study with clear description of the methods. I think the use of a clinical endpoint is excellent.

REVIEWER	Ebbe Thinggaard Post. Doc Copenhagen Academy for Medical Education and Simulation (CAMES) Blegdamvej 9 2100 Copenhagen Denmark Resident Dept. of Obstetrics and Gynaecology University Hospital Hvidovre Kettegård alle 30 2650 Hvidovre Denmark
REVIEW RETURNED	02-Aug-2018

GENERAL COMMENTS	Overall impression: The article is a description of a study protocol where the authors aim to investigate the effects of a comprehensive gamified endoscopy curriculum on doctor's clinical performance. They will conduct a randomized interventional trial and aim to include a total of 36 participants, 18 participants in each arm. The article is well written, the subject is highly relevant and appropriate methods have been chosen to investigate the research question. Abstract:

	In the abstract line 8-9 the authors write: “The application of gamification has not previously been evaluated in the context of a procedural skills simulation-based training curriculum”. This, I think, is an overstatement and should be revised. Something like "No studies that have investigated the effects of a comprehensive gamification curriculum on doctors’ endoscopy skills”, would be more appropriate. The authors use the header: “Article summary: Strengths and limitations of this study:” I would recommend deleting: “Article summary:” and just use “Strengths and limitations”. Methods and Analysis: The rationale for the choice of methods is described in sufficient detail and the assessment tools are appropriate for the intended interpretation of scores. Outcome measures: There seems to be a high number of secondary outcome measures. I would advise that some of these were called exploratory outcome measures instead. I think it is good that the authors have chosen to include several outcome measures as this will provide the authors with a lot of data that may be used for post-hoc explorative research and provide data for better sample size calculations for future interventional studies. Sample size: There seems to be an error here. If you wish to accommodate for a 10% dropout, 19 participants should be included in each group. Use of educational theory: I would like to recommend the authors for their use of appropriate educational theory, studies on simulation training founded on relevant theoretical considerations are highly warranted. Conclusion: I would not use this word as it can be misleading when used in a study protocol. This section should be given a more appropriate header such as; “discussion” or “summary”.

REVIEWER	Dr. Zoë Hoare NWORTH, Bangor University, UK
REVIEW RETURNED	19-Sep-2018

GENERAL COMMENTS	The protocol for the RCT evaluating the gamification of simulation-based endoscopy training appears to be a very interesting concept. However, there are a number of areas where this protocol could be strengthed by giving further details about methods and procedures. I have outlined a number of these areas in my comments below which run in order of the information presented in the manuscript. 1) Randomisation. The procedure used for generating the allocation lists needs to be made explicit - what type of algorithm is being used - are there any stratification variables being used. If participants are being assigned by one investigator (MP) then it cannot be stated in the next sentence that investigators will be blind to allocation. Why are sealed envelopes needed as well as the online system? It is unclear to me whether the recruitment and randomisation will run sequentially over a period of time or will be a number of 'group' randomisations. It would be useful to expand upon the training course that this is running in parallel with to allow an understanding

	of the context this is being developed under (training course is only mentioned in the appendices) 2) Gamification cirriculum - A unique ID is given to 'hide' the identity of the indivudals but they can track their own performance but this is undone somewhat by the use of the 'wearable' medallion which notably identifies the participant with the highest overall ranking at the end of each hour of practice. 2) Consent - The primary outcome is noted as the clinical performance on two 'live' endoscopies. Only by finding the consent form in the appendices did it become clear to me that the patients for these 'live' procedures were being asked to give permission. 3) Primary outcome - this is stated as the 'clinical performance during two live colonoscopies 4 to 6 weeks after training'. Is this one procedure at 4 weeks and one at 6 weeks or two on the same mixture of the two? This is important as this could have an impact on the analysis proposed and the conclusions that can be drawn from this. How this outcome is defined and measured needs to be clearly defined. 4) Analysis Plan - Mixed factor ANOVAs are being proposed where pre-test scores are being included as outcomes scores. By formulating the model in this way the question being answered is slightly different to considering the models by adjusting for the pre-test values. Any pre-test differences in the groups will not be attributable to the intervention but rather chance from the randomisation process. Would it not be better to consider pre-test scores as co-variates rather than using them in this fashion in the repeated measures model? See J.Twisk et al. “Different Ways to Estimate Treatment Effects in Randomised Controlled Trials.” Contemporary Clinical Trials Communications 10 (2018): 80–85. The primary outcome is being assessed using a 2 group 2 measurement model - it is not clear to me why there are two measurements being used of live procedures and when they are measured relative to one another - is the hypothesis that these measurements will not demonstrate any difference? The implication of the design is that the gamification curriculum works for cohorts of trainees rather than sequential trainees (it isn't clear how this is linked to the recruitment process), but one concern I would have that the analysis plan (and sample size) would not have ability to pick up on is the potentially demoralisation of those trainees who fail to score well under the gamification curriculum. A mediation analysis is being proposed using path-analytic framework - this could possibly be quite restrictive given the sample size proposed. There are a huge number of analyses being performed but there is no acknowledgement of the impact of multiple testing on the results, with indication of 5) Sample size estimation - The sample size estimation is based on a very large hypothesised effect size with a 2 group, 3 measurement design - however the primary analysis is based on a 2 group 2 measurement design indicating that the power of the study is not properly linked to the primary outcome analysis. Power of 80% is

	generally lower than what would be expected for rigorously designed studies. It would be useful to expand that effect size into the relative impact on the clinical measure - what difference does this have on the JAG DOPS? This will allow the reader to understand the 6) Feasibility - the dates suggest that data collection is already complete on this study? 7) Appendices – SPIRIT checklist is provided but from what I can see not completed. I'm not sure of the relevance of including all study documentation as appendices for the publication. This makes the paper very unwieldy and as some of these are standardised measures I'm not sure they should be re-produced in this format. However reading some of these appendices has given the additional information for earlier areas but it is unlikely the general reader will look at this additional detail.

VERSION 1 – AUTHOR RESPONSE

REVIEWER #1:
1. Firstly, when did the inclusion of participants start?
Participant inclusion started on June 2017. We have added this to the Methods section (pg. 3).

2. Secondly how will the authors approach a potential sex related difference in performance? Former studies in simulation-based training indicates a performance gap between women and men favoring the latter (Surg Endosc. 2015 Nov;29(11):3065-73. doi: 10.1007/s00464-015-4092-2. Epub 2015 Jan 29. "Gender differences in the acquisition of surgical skills: a systematic review." Ali A1, Subhi Y2,3, Ringsted C4,5, Konge L2.)
Thank you for this suggestion. We have added a sensitivity analysis with the gender covariate for our primary outcome in the Methods section (p. 7).
REVIEWER #2:
NA
REVIEWER #3:
1. In the abstract line 8-9 the authors write: "The application of gamification has not previously been evaluated in the context of a procedural skills simulation-based training curriculum." This, I think, is an overstatement and should be revised. Something like "No studies that have investigated the effects of a comprehensive gamification curriculum on doctors' endoscopy skills", would be more appropriate.
We have changed this sentence accordingly.

2. The authors use the header: "Article summary: Strengths and limitations of this study:" I would recommend deleting: "Article summary:" and just use "Strengths and limitations."
We have deleted this phrase.

3. Outcome measures:
There seems to be a high number of secondary outcome measures. I would advise that some of these were called exploratory outcome measures instead. I think it is good that the authors have chosen to include several outcome measures as this will provide the authors with a lot of data that may be used for post-hoc explorative research and provide data for better sample size calculations for future interventional studies.
We have changed the last three Secondary outcome measures to "Exploratory outcome measures" (pg. 6).

4. Sample size: There seems to be an error here. If you wish to accommodate for a 10% dropout, 19

participants should be included in each group.
Thank you for pointing out this error – we have adjusted it to reflect a 5% dropout rate instead (i.e. 36 total participants) (pg. 9).

5. Conclusion: I would not use this word as it can be misleading when used in a study protocol. This section should be given a more appropriate header such as; “discussion” or “summary.”
We have changed this phrase to “Discussion” as suggested.
REVIEWER #4:
1. Randomisation. The procedure used for generating the allocation lists needs to be made explicit - what type of algorithm is being used - are there any stratification variables being used. If participants are being assigned by one investigator (MP) then it cannot be stated in the next sentence that investigators will be blind to allocation. Why are sealed envelopes needed as well as the online system? It is unclear to me whether the recruitment and randomisation will run sequentially over a period of time or will be a number of 'group' randomisations. It would be useful to expand upon the training course that this is running in parallel with to allow an understanding of the context this is being developed under (training course is only mentioned in the appendices)
We generated a random sequence of numbers using an online sequence generator. One author placed cards labelled with numbers into sealed envelopes and delivered them to another author. A second author, who did not see the allocation sequence, gave these envelopes out to participants as they arrived for the course. The first author who generated the sequence was not present when envelopes were handed out. In this fashion, investigators were blinded to group allocation. We have clarified this in the section “Methods and Analysis: Experimental Design: Training Intervention.” The study takes place during delivery of a training course (“University of Toronto Endoscopic Simulation Course.” This course is delivered to entry level 4th and 5th year gastroenterology residency trainees. Recruitment for this study is planned to take place primarily from participants of this course.

2. Gamification curriculum - A unique ID is given to 'hide' the identity of the individuals, but they can track their own performance, but this is undone somewhat by the use of the 'wearable' medallion which notably identifies the participant with the highest overall ranking at the end of each hour of practice.
We thank the reviewer for pointing this out. It is indeed true that the wearable medallion identifies the participant with the highest ranking. We have added this limitation into the discussion section.

3. Consent - The primary outcome is noted as the clinical performance on two 'live' endoscopies. Only by finding the consent form in the appendices did it become clear to me that the patients for these 'live' procedures were being asked to give permission.
We have added in a sentence into the section “Methods and Analysis: Experimental Design: Delayed Testing.”

4. Primary outcome - this is stated as the 'clinical performance during two live colonoscopies 4 to 6 weeks after training'. Is this one procedure at 4 weeks and one at 6 weeks or two on the same mixture of the two? This is important as this could have an impact on the analysis proposed and the conclusions that can be drawn from this. How this outcome is defined and measured needs to be clearly defined.
The procedures were conducted on the same day, one after another, on a day that was between 4 and 6 weeks after completion of the course. We have clarified this in the “Methods and Analysis: Experimental Design: Delayed Testing.”

5. Analysis Plan - Mixed factor ANOVAs are being proposed where pre-test scores are being included as outcomes scores. By formulating the model in this way, the question being answered is slightly different to considering the models by adjusting for the pre-test values. Any pre-test differences in the groups will not be attributable to the intervention but rather chance from the randomisation process.

Would it not be better to consider pre-test scores as co-variates rather than using them in this fashion in the repeated measures model? See J.Twisk et al. “Different Ways to Estimate Treatment Effects in Randomised Controlled Trials.” Contemporary Clinical Trials Communications 10 (2018): 80–85.
We have reviewed the attached paper. Thank you for providing us with an informative guide on the matter. We, however, chose to keep the mixed factor ANOVA model as this is consistent with our previous studies published on the topic, which will allow for model comparison.

6. The primary outcome is being assessed using a 2 group 2 measurement model - it is not clear to me why there are two measurements being used of live procedures and when they are measured relative to one another - is the hypothesis that these measurements will not demonstrate any difference?
The procedures were conducted on the same day, one after another, on a day that was between 4 and 6 weeks after completion of the course. We have clarified this in the “Methods and Analysis: Experimental Design: Delayed Testing.”

7. The implication of the design is that the gamification curriculum works for cohorts of trainees rather than sequential trainees (it isn't clear how this is linked to the recruitment process), but one concern I would have that the analysis plan (and sample size) would not have ability to pick up on is the potentially demoralisation of those trainees who fail to score well under the gamification curriculum.
Although considered in the planning of the study, we have not observed any signs of demoralization in our current sample. Furthermore, despite the anecdotal nature of this evidence, we do not believe that demoralization will be observed, as participants’ performance is relative to their respective cohort. That being said, it is possible that participants scoring in the lower tier of their cohort could experience some frustration and we not have built this into the analysis – we have addressed this as a potential limitation in the “Discussion” (pg. 8).

8. A mediation analysis is being proposed using path-analytic framework - this could possibly be quite restrictive given the sample size proposed.
We have removed the proposed analysis.

9. There are a huge number of analyses being performed but there is no acknowledgement of the impact of multiple testing on the results, with indication of
We have added a statement indicating that this consideration will be built into the analyses in the Results section (pg. 6).

10. Sample size estimation - The sample size estimation is based on a very large hypothesised effect size with a 2 group, 3 measurement design - however the primary analysis is based on a 2 group 2 measurement design indicating that the power of the study is not properly linked to the primary outcome analysis. Power of 80% is generally lower than what would be expected for rigorously designed studies. It would be useful to expand that effect size into the relative impact on the clinical measure - what difference does this have on the JAG DOPS? This will allow the reader to understand the
We have modified the sample size estimation in light of this suggestion, by changing the primary analysis to be based on a 2 group 2 measurement design in the Methods section and have updated our estimate accordingly (p. 7). Regarding the expansion of the effect size into the relative impact on the clinical measure, however, we are unsure what is being referred to by the reviewer.

11. Feasibility - the dates suggest that data collection is already complete on this study?
At the time this study was submitted, primary data collection was not complete. Currently, participants have completed the study. Their performances, however, have not been fully assessed by expert endoscopists. We have changed the tense in the manuscript to reflect past tense as appropriate.

12. Appendices – SPIRIT checklist is provided but from what I can see not completed. I'm not sure of the relevance of including all study documentation as appendices for the publication. This makes the paper very unwieldy and as some of these are standardised measures I'm not sure they should be re-produced in this format. However. reading some of these appendices has given the additional information for earlier areas but it is unlikely the general reader will look at this additional detail.
Thank you for catching this error. We have now attached the correct version of the SPIRIT checklist. With respect to additional appendices, we believe it is important to have this information available should readers have further questions about the methodology. Additionally, the appendices provide sufficient detail to allow for a reproduction of this trial.

VERSION 2 – REVIEW

REVIEWER	Ebbe Thinggaard, MD, PhD Copenhagen Academy for Medical Education and Simulation Department of Obstetrics and Gynaecology, Hvidovre University Hospital
REVIEW RETURNED	21-Nov-2018

GENERAL COMMENTS	Congratulations on a well written article. I look forward to seeing the results of the study.

REVIEWER	Dr Zoë Hoare NWORTH CTU, Bangor University Wales UK
REVIEW RETURNED	30-Nov-2018

GENERAL COMMENTS	Majority of the comments have been satisfactorily. Apologies for one of my comments having been cut off; what I meant by this comment was whether the effect size indicated for the JAG DOPS could be quantified in terms of a X point change of the measure making it more relatable for the reader who doesn't think in effect sizes. I accept this make not make sense for all measures and admit I'm not familiar with this particular outcome. I'm not convinced that comparability if analysis with previously used methodology is an appropriate reason for choice of analysis model. However I'll defer that decision to the editor as this is a protocol paper and changes to the planned analysis methods can occur beyond this point. I appreciate the limitations added to the end of the paper but these could be better phrased for clarity to the reader

VERSION 2 – AUTHOR RESPONSE

REVIEWER #4:
1. Majority of the comments have been satisfactorily. Apologies for one of my comments having been cut off; what I meant by this comment was whether the effect size indicated for the JAG DOPS could be quantified in terms of a X point change of the measure making it more relatable for the reader who doesn't think in effect sizes. I accept this make not make sense for all measures and admit I'm not familiar with this particular outcome.
• We appreciate that not all readers of this protocol may be familiar with this outcome measure. However, the target audience for this article, academic endoscopists and gastroenterologists, are

largely familiar with JAG DOPS and will be familiar with the effect size provided. Additionally, for educational assessment measures it is not typical to quantify effect size in terms of a “X point change” as suggested.

2. I’m not convinced that comparability if analysis with previously used methodology is an appropriate reason for choice of analysis model. However, I’ll defer that decision to the editor as this is a protocol paper and changes to the planned analysis methods can occur beyond this point.

- We chose the current analyses to ensure continuity with our previous studies in the area. Furthermore, those previous studies have undergone peer-review before publication, which lends credibility to the included analyses.

3. I appreciate the limitations added to the end of the paper, but these could be better phrased for clarity to the reader

- We have rephrased the limitations to make them clearer for the reader.