PEER REVIEW HISTORY

BMJ Paediatrics Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Rater training for standardized assessment of Objective Structured Clinical Exams in rural Tanzania
AUTHORS	Sigalet, Elaine L Matovelo, Dismas Brenner, Jennifer L Boniphace, Maendeleo Ndaboine, Edgar Mwaikasu, Lusako Shabani, Girles Kabiligi, Julieth Mannerfeldt, Jaelene Singhal, Nalini

VERSION 1 – REVIEW

REVIEWER	Reviewer name: Dr. Elizabeth Petersen Institution and Country: Brigham and Women's Hospital, United States Competing interests: None
REVIEW RETURNED	13-Sep-2020

GENERAL COMMENTS	The purpose of this study was to describe a 3-day OSCE rater training curriculum in rural Tanzania. The concept for the study is important for improving the outcomes for HBB, ECEB and BAB training programs in LMIC. Abstract - See comments in “Methods” section below for recommendations on the addition of details regarding how the raters were trained. - Conclusion- The results of your study show moderate agreement between 6 raters for standardized case scenarios and that raters were only able to identify average level of proficiency (which is the level of pass) 50% of the time, or equal to chance. Please comment on how you concluded that the rater training curriculum was effective. Background - Please condense the first 3 paragraphs of the background section to 2 paragraphs as able. Methods - Please describe in more detail the 3-day rater training so that it could be replicated in another setting. o You might consider referencing the scoring checklists for the HBB, ECEB and BAB OSCEs in your manuscript. o Did the raters receive specific training on how to facilitate these OSCEs and/or interpret the OSCE checklists prior to scoring the standardized learners? If so, please describe what this training entailed as it is essential for replication.

	o Did the raters receive any training on potential sources of error and/or bias in scoring learners and how to avoid these common pitfalls? Please comment. o How were the standardized learners prepared for each case scenario to ensure uniformity? Did each rater watch the exact same scenario at the same time? Please comment. o How did the standardized case scenarios differ to reflect learners functioning at poor, acceptable and excellent proficiency levels? Please comment. o Figure 1: Please comment on how the case scenarios scored were distributed over the 3-day course. Were OSCEs for either HBB, ECEB and BAB used on different days? Were the same scenarios revisited on subsequent days for repetition? o Setting- Here you indicate that the study was conducted over 2-days, but elsewhere in the paper you report that this was a 3-day rater training. Please clarify. Results - Do you have data on interrater reliability for in country raters prior to undergoing rater training? This would provide context for whether your training represents an improvement from prior. - One of the most notable results appears to be the fact that raters only identified average proficiency (or the level to pass the OSCE) 50% of the time, which is equal to chance. Please see discussion section regarding need for comments on this. - Please attempt to condense the field notes section to highlight general themes, as the specific details are not as pertinent to the outcomes of your study. Discussion - I agree with the authors that in country rater training is important for the successful administration of OSCEs to assess proficiency of learners taking HBB, ECEM and BAB courses. However, given that the results demonstrate that raters were only able to identify average proficiency successfully 50% of the time can you conclude that this rater training was successful and should be replicated? - Does achievement of moderate rater agreement matter if the raters are only able to identify average proficiency 50% of the time? - Please comment on whether you feel that the number of raters (n=6) is a limitation in this study, especially as it pertains to the assessment of interrater reliability. - The raters in this study were chosen as a result of their prior success in these courses? Is this a potential source of bias this this study? Does it affect the generalizability? Please comment. - How might a rater training program be setup to help raters discriminate between learners with borderline performance on the OSCE? Conclusion - Please provide comments as to why you believe that this rater training program was effective. Figure 2 - The right side of Figure 2 is cut off in my view of the manuscript. General comments - Please review manuscript for spelling and grammatical errors.

	- Citations within the manuscript are noted with both [ ] and as superscripts. Please unify.

REVIEWER	Reviewer name: Dr. Christiana M Russ Institution and Country: Children's Hospital Boston, Massachusetts, United States Competing interests: None
REVIEW RETURNED	15-Sep-2020

GENERAL COMMENTS	Overall this is a well designed, and well described study, that reminds us that the devil is in the details for standardizing and scaling educational interventions, particularly complex clinical skills. I think the relatively mixed findings are important to report. While I agree with the conclusion that training of in-country raters is feasible I am not clear that I agree entirely on the efficacy, demonstrating only moderate agreement among raters (and a decrease in agreement for BAB). I think the authors could provide more context as to what is considered 'good enough' for validation of these scales in other contexts. And I think the fall-off of agreement for BAB needs greater consideration in the discussion. What was it about that rating system or those OSCEs that made it different from HBB or ECEB? I also noted that the first and last author for this study are both from the HIC partner. I wonder given that this work was done entirely in Tanzania if there is a way for a Tanzanian partner to get more authorship credit in some way. Background: More information on what is acceptable inter-rater variation and what has been achieved would be helpful for better context. Are the OSCE resources provided by these standardized training already validated in some way? Method: P 8 paragraph 1 - Participant selection - How long since the participants had undergone formal training in each of these programs? There is good data about fall-off of proficiency over time, so this could be quite relevant particularly without a refresher. Both figures are quite helpful in clarifying the iterative method used. Please describe more clearly exactly when the data were collected that was used for the study comparison. Were any ratings done prior to formal training beginning? Both figures could use some more formatting, and were quite small. Figure 1 was a bit blurred so difficult to read. For Figure 2 the text on the right side margin appears to be cut off. You describe in the Discussion adding sub-item tracking boxes to help clarify multi-step OSCE items. At what point did you add those in? Do you have data before and after doing this and did it improve results? Discussion: I recommend reorganizing the discussion section to first set these results in better context up front. You note clearly at the end of the first paragraph that these challenges are not specific to LMIC, so perhaps discussion about the more global challenges with OSCE validation, and similar inter-rater values obtained in similar (paragraph 3) and also different contexts would be helpful. I am

	left curious as to what inter-rater values are considered appropriate in high-stakes OSCES in more resourced settings. P 13 line 13-22 – discussion about overestimation and underestimation is confusing. Perhaps use one sentence for each type of risk. P 13 line 31 – I would avoid the term 'developed' and perhaps refer to HIC or highly resourced. I also am not clear what a 'global health rating scale' is. More discussion is needed as to why the kappa for the BAB OSCE rating went down. You have globally reported 4 challenges: variation in agreement on standard, multi-step items, not understanding terminology and not counting actions that were not verbalized. Were there more of these challenges in the BAB OSCE items? Were there specific BAB items that were more difficult to get agreement upon? I also think you should clarify why there remained significant disagreement at the end. Did differences of opinion persist over the standard? It seems these other challenges could be overcome with your iterative process. Conclusion – I am not clear in the end that you can say this intervention was entirely effective. Feasible, yes. Important, yes – and indeed the lack of efficacy shows a gap that really needs to be addressed. From these data it appears it was somewhat effective for HBB, significantly effective for ECEB, and not effective for BAB. There is good learning in there, but more study is indeed warranted!

REVIEWER	Reviewer name: Dr. Peter Flom Institution and Country: Peter Flom Consulting, United States Competing interests: None
REVIEW RETURNED	19-Sep-2020

GENERAL COMMENTS	I confine my remarks to statistical aspects of this paper. A general comment is that it is probably not a good idea to collapse categories in the way the authors did. It would be stronger, statistically, to leave all the categories as is and use Kendall tau coefficients. More specific comments p. 4 line 24 "vulnerable" seems like the wrong word here. Vulnerable to what? Remove p values and significance tests throughout. A test that kappa is not 0 is not useful. Of course it's not 0. The question is how big it is. The authors could include the standard errors of the kappa as an indicator of how good the estimate of kappa is. Table 2 - I don't understand what the kappa statistics for individual levels mean. (e.g. 0.32, 0.32, 0.63. 0.46 etc. Peter Flom

VERSION 1 – AUTHOR RESPONSE

Dear Editors for BMJ Paediatrics Open Re Manuscript #2020-000856, "Rater training for standardized assessment of Objective structured Clinical exams in rural Tanzania,"

We are grateful for the opportunity to revise the manuscript, as we too believe that the concept of rater training is important in all settings.

Responses to Comments of Reviewer #1.

Abstract
See comments in "Methods" section below for recommendations on the addition of details regarding how the raters were trained.

We amended the sentence under Method section of abstract on page 4 to read:

Researchers used Zabar's criteria to critique rater agreement and mitigate measurement error during score review. Descriptive statistics, Fleiss' kappa and field notes were used to describe results.

Conclusion- The results of your study show moderate agreement between 6 raters for standardized case scenarios and that raters were only able to identify average level of proficiency (which is the level of pass) 50% of the time, or equal to chance. Please comment on how you concluded that the rater training curriculum was effective.

We agree interpreting the efficacy of our results is challenging. To increase clarity, we changed the first line in the conclusion stating: Our study shows in rural Tanzania training of in country raters is feasible and effective to read: (abstract Page 5)

Our study shows ***training of in-country raters resulted in the discernment of acceptable proficiency 50% of the time, despite moderate rater agreement****. Rater training is critical to ensure that the potential of training programs translates to improved outcomes for mothers and babies; more research into the concepts and training for discernment of competence in this setting is necessary.*

In our conclusion we also made changes to this statement and the conclusion of the main body of the manuscript page 17/18 first paragraph. We infer that our training was effective in terms of learning over the three days because we saw an increase in rater agreement over HBB and ECEB. The falloff in BAB may be attributable to fatigue as those scenarios were role played last on day 3. We would point out that there is very limited literature

examining these rater reliabilities in any setting and this speaks to the need for further studies.

Our result show that rater training in an LMIC setting is critical for administering OSCE based learner assessments. Clinician everywhere need ongoing training, but to optimize learning and then translate this to improved outcomes for mothers and babies, this training must be informed by truly objective evaluations . ***Our study shows in rural, Tanzania, training of in-country raters is possible and can lead to an IRR which is similar to previous studies. Improved standardization and attention to the relationships between IRR and the accurate discernment of participant performance would provide insight into needed modifications, which in turn may lead to greater accuracy in rating competence.*** *More research is warranted.*

Background

- Please condense the first 3 paragraphs of the background section to 2 paragraphs as able.

We have done this – See page 6 and 7

Methods

- Please describe in more detail the 3-day rater training so that it could be replicated in another setting.

To add clarity, we added the following sentences under the rater training curriculum page 9. We have revised figure 2 so it is readable on the page.

Checklists were reviewed prior to scoring practice Day 1 of training to ensure raters were familiar with OSCE items and how to use the checklist in scoring. Raters observed a scenario, with a predetermined level of proficiency. Training of raters occurred in the score review, with faculty leading discussions to discern the underlying ideas or concepts which may have led to the disagreement. Raters learned about potential sources of error in the discussion of rater disagreements in score review. Faculty discussed the importance of mitigating these sources of error to improve score reliability. Scenarios with disagreement on two or more items were repeated.

You might consider referencing the scoring checklists for the HBB, ECEB and BAB OSCEs in your manuscript.

These are referenced under Evaluation tools under the Method section. They were drawn from the training materials [1-4]

"The OSCEs used were drawn from training program materials.[1-4]"

Did the raters receive specific training on how to facilitate these OSCEs and/or interpret the OSCE checklists prior to scoring the standardized learners? If so, please describe what this training entailed as it is essential for replication.

We did not focus on training raters as facilitators. WE did review the checklists with them prior to the first mock scoring practice for each OSCE. Review of the checklist items was ongoing as we incurred challenges related to the items.

We added the following sentence under Method section, subheading The Rater Curriculum Page 9- to provide more clarity about rater familiarity with the checklists

Checklists were reviewed prior to scoring practice Day 1 of training to ensure raters were familiar with OSCE items and how to use the checklist in scoring.

Did the raters receive any training on potential sources of error and/or bias in scoring learners and how to avoid these common pitfalls? Please comment.

We added the following sentences under the method section, subheading rater training curriculum on page 9-10 to clarify

Training of raters occurred in the score review, with faculty leading discussions to discern the underlying ideas or concepts which may have led to the disagreement. Raters learned about potential sources of error in the discussion of rater disagreements in score review. Faculty discussed the importance of mitigating these sources of error to improve score reliability. Scenarios with disagreement on two or more items were repeated.

o How were the standardized learners prepared for each case scenario to ensure uniformity? Did each rater watch the exact same scenario at the same time? Please comment.

Each rater watched the exact same scenario at the same time – to **add clarity we added the following sentence under the Method Section: subheading Design page 8**

We removed "each participant selected to be a rater independently scored each scenario" and added

All six raters observed and scored the exact same scenario at the same time, making judgements about observed behaviors independent of discussion with each other.

Under the Method section: evaluation tools page 9 we added:

To standardize the proficiency level deemed to be acceptable in a scenario, a priori the researchers used the clinical consequences of an action to inform the scoring, which was then used to plan the actions role played in the scenario.

o How did the standardized case scenarios differ to reflect learners functioning at poor, acceptable and excellent proficiency levels? Please comment.

We have addressed this in the above comment.

Figure 1: Please comment on how the case scenarios scored were distributed over the 3-day course. Were OSCEs for either HBB, ECEB and BAB used on different days? Were the same scenarios revisited on subsequent days for repetition?

To add clarity, we added scenario distribution to each day. This is reinforced in Table 2 and 3 where we provide a day by day breakdown of the results of proficiency levels and number of scenarios by course and by day.

o Setting- Here you indicate that the study was conducted over 2-days, but elsewhere in the paper you report that this was a 3-day rater training. Please clarify.

We corrected this in the setting section under Method page 7:

The study was conducted in Kwimba District located in Mwanza Region, Tanzania over three days; two days in April 2018 and one day in May 2018.

Results

- Do you have data on interrater reliability for in country raters prior to undergoing rater training? This would provide context for whether your training represents an improvement from prior.

Unfortunately, the answer is No. The raters had never rated before. We would hope to see a repetitive type of study that could look at this again in the future.

- One of the most notable results appears to be the fact that raters only identified average proficiency (or the level to pass the OSCE) 50% of the time, which is equal to chance. Please see discussion section regarding need for comments on this.

We have addressed this in the discussion.

- Please attempt to condense the field notes section to highlight general themes, as the specific details are not as pertinent to the outcomes of your study.

We removed the specifics of the challenges with field notes in the main body of the manuscript as these are well described in Table 4 and modified Table 4 using themes.

Discussion

- I agree with the authors that in country rater training is important for the successful administration of OSCEs to assess proficiency of learners taking HBB, ECEM and BAB courses. However, given that the results demonstrate that raters were only able to identify average proficiency successfully 50% of the time can you conclude that this rater training was successful and should be replicated?

Our intention was to show efficacy in relation to comparable studies and based on improvement over training time, but we appreciate from a reader perspective this may not be clear. To improve clarity, we added further detail to the methodology on page 14. Which describes our use of an inquiry model to debrief participants. This was the fundamental unit for learning. Each rater attended 42 debriefing over the 3-day course.

We make explicit the result of moderate IRR yet only approximately 50% discernment of average to excellent which is a pass in these courses. Furthermore, in literature IRR is often reported as a proxy for validation. This is problematic in any setting. Actual reporting of rater reliability is rarely done and so we cannot say definitely is moderate IRR and 50% discernment are a successful outcome of training. We would point to the fact that there was improvement in rater agreement (Table 1) over the 3 days for HBB and ECEB which infers learning. There are no benchmarks on which to decide whether this type of training is adequate. However, our results underline the importance of this type of training; the lack of benchmarks points out the critical importance of this work and the need for it to be improved upon in future iterations.

- Does achievement of moderate rater agreement matter if the raters are only able to identify average proficiency 50% of the time?

See above.

- Please comment on whether you feel that the number of raters (n=6) is a limitation in this study, especially as it pertains to the assessment of interrater reliability.

We would agree that in an ideal setting more raters would provide more insight however in a LMIC setting, the logistics involved in recruiting such participants meant for that 6 was as many as we could recruit.

- The raters in this study were chosen as a result of their prior success in these courses? Is this a potential source of bias this this study? Does it affect the generalizability? Please comment.

Yes they were chosen as a result of their success in previous courses however it is an educational fundamental that to assess someone you have to have competence in the field so the pool of candidates is necessarily limited to those who have exhibited competency previous in the content area.

Random selection of participants to be assessors was not possible in this setting.

How might a rater training program be setup to help raters discriminate between learners with borderline performance on the OSCE?

A training program to help raters discriminate between borderline and acceptable competence would need significant investment in understanding the rater's knowledge and abilities to provide objective scores based on observed behaviors. This would then provide a useful framework for identifying rater's knowledge and performance gaps, which could then inform curriculum redesign. In this context the advantages of such an investment would need to be weighed against other options such as introduction of a global rating scale. We feel the latter strategy is more likely to be effective consistent with worldwide use. This has a cost advantage and implementation is much simpler.

Conclusion
- Please provide comments as to why you believe that this rater training program was effective.

We have addressed this throughout from the previous reviewer comments.

Figure 2
- The right side of Figure 2 is cut off in my view of the manuscript.

We have revised figure 2.

General comments
- Please review manuscript for spelling and grammatical errors.
- Citations within the manuscript are noted with both [ ] and as superscripts. Please unify.

We have addressed these.

Reviewer: 2
Reviewer: 2

Comments to the Author
Overall this is a well-designed, and well described study, that reminds us that the devil is in the details for standardizing and scaling educational interventions, particularly complex clinical skills. I think the relatively mixed findings are important to report.
While I agree with the conclusion that training of in-country raters is feasible I am not clear that I agree entirely on the efficacy, demonstrating only moderate agreement among raters

(and a decrease in agreement for BAB). I think the authors could provide more context as to what is considered 'good enough' for validation of these scales in other contexts.

We agree this has been addressed in our discussion(page 14-16) and abstract (page 5) and manuscript conclusion (page 17). The reviewers comment as to what is good enough for validation of rating scale is a common problem not isolated to LMICs. Unfortunately, rater discrimination is rarely reported, and the IRR is typically used as a proxy. We believe we saw an improvement in IRR, and we would suggest using an additional data point of a global rating scale to help improve the objectivity involved in scoring.

And I think the fall-off of agreement for BAB needs greater consideration in the discussion. What was it about that rating system or those OSCEs that made it different from HBB or ECEB?

This is a fair comment. Our methodology did not change with this OSCE; we speculate that rater fatigue may have played a role. We added the following statement to make this explicit on page 16, paragraph 3 of the discussion.

The fall-off in rater agreement for BAB Day 3 was unexpected but may be in part related to the timing of these scenarios' day 3; they were the last role plays of the day and rater fatigue may have played a role.

I also noted that the first and last author for this study are both from the HIC partner. I wonder given that this work was done entirely in Tanzania if there is a way for a Tanzanian partner to get more authorship credit in some way.

Much discussion went into this decision. Authorship followed ICMJE guidelines. This was our first work together, however our local leader Dismas Matovelo will be first author on two upcoming manuscripts.

Background:
More information on what is acceptable inter-rater variation and what has been achieved would be helpful for better context. Are the OSCE resources provided by these standardized training already validated in some way?

The standardized training has been extremely rigorously established and endorsed by Jhpeigo and American Academy of Pediatrics. Course materials provide guidance with minimizing sources of measurement error in the set up of the OSCE but do not speak to a training rater curriculum.
Previous studies with the exception of one [15] examining the knowledge transfer imparted by the course have always used an external partner as part of the OSCE evaluation.

We speak to this in the introduction paragraph two on page 6 and in the discussion pages 14-15.

Method:

P 8 paragraph 1 - Participant selection - How long since the participants had undergone formal training in each of these programs? There is good data about fall-off of proficiency over time, so this could be quite relevant particularly without a refresher.

A great comment and point. Raters had all participated in these training programs in the last year. To add clarity, we supplemented the sentence on page 7

Selection was based on their demonstrated proficiency in previous Newborn Maternal training workshops conducted in the previous year. All trainees were clinically active in their health facility settings.

Both figures are quite helpful in clarifying the iterative method used. Please describe more clearly exactly when the data were collected that was used for the study comparison. Were any ratings done prior to formal training beginning?

To add clarity, we added the following sentence under the Method, subheading Design page 8

Raters attended rater training prior to any formal scoring of workshop participants. …. Scores were collected and then reviewed with the raters; areas of disagreement were explored, using an inquiry approach for debriefing.

We also added data collection in figure 1 to make it explicit.

Both figures could use some more formatting and were quite small. Figure 1 was a bit blurred so difficult to read. For Figure 2 the text on the right side margin appears to be cut off.

Figures are reformatted and uploaded.

You describe in the Discussion adding sub-item tracking boxes to help clarify multi-step OSCE items. At what point did you add those in? Do you have data before and after doing this and did it improve results?

We modified the following statement to add clarity under the discussion section, paragraph 5 page 16:

To address this gap, we added sub-item tracking boxes when this challenge was identified Day 1; the use of this strategy warrants further study.

Unfortunately, we did not track before and after to provide information on this but acknowledge that would have been helpful.

Discussion:
I recommend reorganizing the discussion section to first set these results in better context up front. You note clearly at the end of the first paragraph that these challenges are not specific to LMIC, so perhaps discussion about the more global challenges with OSCE validation, and similar inter-rater values obtained in similar (paragraph 3) and also different contexts would be helpful. I am left curious as to what inter-rater values are considered appropriate in high-stakes OSCES in more resourced settings.

We agree this is problematic in all settings and have brought the general discussion forward in paragraph 1. We address the general subject of IRR in paragraph 2. Our focus was to compare this to similar studies with these courses where reports are rare. In a high stakes OSCE's medical education experts would suggest a "very good" level of reliability to support assessment of competence (Kappa >0.81). However, except for a very few high stakes exams like ACLS, reliability is rarely reported.

P 13 line 13-22 – discussion about overestimation and underestimation is confusing.

Perhaps use one sentence for each type of risk.

We agree and were confused too on rereading this. To hope the following is easier to follow: discussion section, paragraph 1 page 14

With overestimation of competence, training programs may have passed clinicians who may need more training to provide safe care on the frontline. The problems of accurate discrimination of competency also affect resource utilization: with underestimation of competence, training programs may be directing the limited resources to clinicians who do not need extra training.

P 13 line 31 – I would avoid the term 'developed' and perhaps refer to HIC or highly resourced.

We have changed developed to HIC as suggested.

I also am not clear what a 'global health rating scale' is.

We may have confused the reader by writing global health rating scale when we should have written global rating scale- so we omitted the word health under the Discussion Section paragraph 2 and defined and referenced it in same section- page 14

A global rating scale allows the rater to evaluate how well a learner performs on a scale of 1 to 5, with 5 reflecting the highest level of competence[28]

More discussion is needed as to why the kappa for the BAB OSCE rating went down. You have globally reported 4 challenges: variation in agreement on standard, multi-step items, not understanding terminology and not counting actions that were not verbalized. Were there more of these challenges in the BAB OSCE items? Were there specific BAB items that were more difficult to get agreement upon?

This was mentioned by reviewer 1 and unfortunately the only difference we noted was the timing of score review practice contexts for BAB day 3. We had completed HBB and ECEB first and BAB was practiced last. We addressed this by adding this context under Discussion paragraph 3 page 15 and 16

The fall-off in rater agreement for BAB Day 3 was unexpected but may be in part related to the timing of these scenarios' day 3; they were the last role plays of the day and rater fatigue may have played a role.

I also think you should clarify why there remained significant disagreement at the end. Did differences of opinion persist over the standard? It seems these other challenges could be overcome with your iterative process.

We are unclear about what the reviewer is referring to. The disagreement mentioned was between the ratings given by the raters and the actual level of proficiency demonstrated by the research team lead. The discussion regarding the difference in opinion formed the basis for the ongoing iterative learning process which included focus score debriefing using inquiry and direct feedback followed by repetitive practice.

Conclusion – I am not clear in the end that you can say this intervention was entirely effective. Feasible, yes. Important, yes – and indeed the lack of efficacy shows a gap that really needs to be addressed. From these data it appears it was somewhat effective for HBB, significantly effective for ECEB, and not effective for BAB. There is good learning in there, but more study is indeed warranted!

We agree and have changed our abstract and manuscript conclusion to reflect this.

Reviewer: 3

Comments to the Author
I confine my remarks to statistical aspects of this paper.

A general comment is that it is probably not a good idea to collapse categories in the way the authors did. It would be stronger, statistically, to leave all the categories as is and use Kendall tau coefficients.

We agree statistically that it would be better to just present all levels of proficiency, but to ensure others can compare in the future we combined the average and excellent categories to align with course guidelines. The third category of excellent was introduced by the in country faculty to explore a more objective methodology for candidate selection for future in country teachers and raters.

More specific comments

p. 4 line 24 "vulnerable" seems like the wrong word here. Vulnerable to what?

We agree – this has been removed.

Remove p values and significance tests throughout. A test that kappa is not 0 is not useful. Of course it's not 0. The question is how big it is. The authors could include the standard errors of the kappa as an indicator of how good the estimate of kappa is.

We agree and removed all references to p values throughout manuscript and tables. We added the standard error as suggested.

Table 2 - I don't understand what the kappa statistics for individual levels mean. (e.g. 0.32, 0.32, 0.63. 0.46 etc.

We agree and see the redundancy here so eliminated these.

Peter Flom

Associate Editor
Comments to the Author:
Please take a close look at the statistical reviewers comments on categorization and how kappa statistics are presented.

We have addressed this by including standard error and removing all references to p values.

Both content reviewers have a number of important comments, most important that only very modest inter rater agreeement is achieved, calling into question whether the entire intervention was even successful. This needs to be addressed in detail.

We may have overstated this claim and have adjusted the language throughout. As we have described in our response to the reviewers, we feel strongly that learning occurred because of the improvement over the three days of training with HBB and ECEB and think that rater fatigue affected the fall off in BAB. We also believe that there is a real need for

sharing our experience so others can understand and address this important element of education and objective assessment in a low resource setting.

Finally, I agree with Dr. Russ' important comment on equity , so please do address the role of the different team members. There is an issue of framing and tone here; even though most of the authors are from Tanzania and the entire project was conducted there, the participants and setting are repeatedly referred to as "in-country" which gives a strong colonial impression.

This was not our intention. The use of the phrase "in country" was the choice of our African partners. In their mind local imparts a stigma.
Much discussion went into the authorship order. Authorship followed ICMJE guidelines. This was our first work together, however our local leader Dismas Matovelo will be first author on two upcoming manuscripts.

Editor in Chief
Discussion 1st sentence delete "This is the first study" see instructions to authors. We do not recommend ever stating "first study", as this is up to others to decide and additionally, you do not know what will be published tomorrow by others. Similarly amend the 2nd sentence in Conclusions.

We removed as suggested.

Be more objective in your conclusions.

We hope we have succeeded in doing this integrating feedback from all reviewers.

Consider showing your results in a Figure. It would attract more readers.

We think the tables are easier to read now after modification based on reviewer feedback.

"Local" may be a better term than "in country" in the text and what this study adds and what is already known sections

Our African partners changed our initial label of "local" to " in country" in revisions as they felt it was a more appropriate way to reference themselves and their context.

We hope this meets with your approval for publication. Many thanks for the opportunity.

VERSION 2 – REVIEW

REVIEWER	Reviewer name: Dr. Christiana M Russ Institution and Country: Children's Hospital Boston, Massachusetts, United States Competing interests: None
REVIEW RETURNED	03-Nov-2020

GENERAL COMMENTS	Overall this is a well designed, and well described study, that reminds us that the devil is in the details for standardizing and scaling educational interventions, particularly complex clinical skills. Thank you for this contribution. I appreciate the revisions that highlight the challenges in these results, and the suggestions for improvement based on these experiences. The edits in methods have significantly improved the clarity. I have only a few minor suggestions in the Discussion section. Discussion: Page 14 line 228 – I believe there is a typo, expecting a period after the word validation. Page 14 line 231 – I would suggest staying away from terms like 'developed world' which some perceive as pejorative, and favor sticking with terms such as LMIC, HIC, etc... Page 15 line 238. – I would remove the last line of this paragraph as it is redundant. Page 15 line 256-257 – I would strike the sentence: 'In a limited resource setting this is challenging to develop and implement...' Video is easier than ever to make, and this is a feasible solution to the challenge you describe of standardizing mock scenarios, with limited personnel. In fact I wonder if these types of videos should be included in course materials - do you think that would have helped in your course? Page 15 Line 260- You noted in the introduction that sources of error can arise from OSCE structure and/or rater objectivity. While rater fatigue is an interesting source for this discrepancy, it might be useful to re-introduce this frame, and comment on whether your team thinks there are any differences in the OSCE structure for BAB compared to HBB and ECEB. From Table 4 it appears that BAB had the greatest number of differing perceptions of practice standards – do you think that played a role?

REVIEWER	Reviewer name: Dr. Peter Flom Institution and Country: Peter Flom Consulting, United States Competing interests: None
REVIEW RETURNED	24-Oct-2020

GENERAL COMMENTS	The authors have addressed my concern and I now recommend publication.

VERSION 2 – AUTHOR RESPONSE

Dear Editors for BMJ Paediatrics Open Re Manuscript #2020-000856, "Rater training for standardized assessment of Objective structured Clinical exams in rural Tanzania,"

Many thanks for the opportunity to publish this important study. We have addressed the suggestions below.

Reviewer: 1

Comments to the Author
The authors have addressed my concern and I now recommend publication.

Reviewer: 2

Comments to the Author
Overall this is a well designed, and well described study, that reminds us that the devil is in the details for standardizing and scaling educational interventions, particularly complex clinical skills. Thank you for this contribution. I appreciate the revisions that highlight the challenges in these results, and the suggestions for improvement based on these experiences. The edits in methods have significantly improved the clarity. I have only a few minor suggestions in the Discussion section.

Discussion:
Page 14 line 228 – I believe there is a typo, expecting a period after the word validation.

We agree and have addressed this. Page 14 Discussion line 229.

Page 14 line 231 – I would suggest staying away from terms like 'developed world' which some perceive as pejorative, and favor sticking with terms such as LMIC, HIC, etc...

We agree and have changed "developed world" to HIC. (line 229)

Page 15 line 238. – I would remove the last line of this paragraph as it is redundant.

We agree and have removed this line

Page 15 line 256-257 – I would strike the sentence: 'In a limited resource setting this is challenging to develop and implement...' Video is easier than ever to make, and this is a feasible solution to the challenge you describe of standardizing mock scenarios, with limited personnel. In fact I wonder if these types of videos should be included in course materials - do you think that would have helped in your coure?

We have removed this line but although we would agree the concept of making a video is easy enough with technology today and we certainly believe that if we had these the training videoss it would have been more successful. However in our experience (Tanzania, Uganda, Malawi and Sierra Leone) – making videos still requires dedicated time, people and more funding: The impact is far reaching: the field needs replacement (cost more money) or the facility works short staff (demand always seem to exceed workforce in these settings), the person attending the video making session must be paid, and lastly they are not trained as actors so it requires a fair amount of mentorship. Maybe the answer lies in getting a dedicated grant just to build video for a rater training curricula- an interesting concept!

Page 15 Line 260- You noted in the introduction that sources of error can arise from OSCE structure and/or rater objectivity. While rater fatigue is an interesting source for this discrepancy, it might be useful to re-introduce this frame, and comment on whether your team thinks there are any differences in the OSCE structure for BAB compared to HBB and ECEB. From Table 4 it appears that BAB had the greatest number of differing perceptions of practice standards – do you think that played a role?

This is a great point so thank you. We added the following frame to make this explicit. (P. 16-

line 262).

Additionally, the greater number of differing perceptions of the practice standard

(Table 4) may have impacted this finding.

Associate Editor
Comments to the Author:
The paper is much improved.

I have a few minor comments:

With the clarifications and revisions, it is clearer that a major contribution of the paper is showing that raters have a hard time with average performance. I think this could be highlighted even more clearly in the abstract, conclusions, and "what this adds" section

We appreciate this perspective and have revised the three areas accordingly.

In the abstract we changed the statement "*Our study shows training of in-country raters resulted in the discernment of acceptable proficiency 50% of the time, despite moderate rater agreement*" to read:
"Our study shows that the in-country raters in this study had a hard time identifying average performance despite moderate rater agreement."

We expanded the following sentence in the conclusion page 17 line 308 to include this:

Our result show that rater training in an LMIC setting is critical for administering OSCE based learner assessments ***especially since the raters in this study had a hard time identifying average performance.***

We modified statement #2 under what this study adds to read:

2. Raters had a hard time identifying average performance despite the achievement of moderate rater agreement.

In the abstract, I suggest not using abbreviations to improve readability

We removed all abbreviations except OSCE as this type of evaluation is known to most as an OSCE.

In the abstract, the sentence beginning, "Our studies shows training…" is the critical sentence, but it isn't clear. I suggest something like "training… of raters… led to moderate interrater agreement and good classification of poor and excellent performance. However, appropriate classification of acceptable or average performance was achieved on 50% of the time."

We addressed this with Reviewer #2 comment:

Our study shows ***that the in-country raters in this study had a hard time identifying average performance despite moderate rater agreement***.

In the methods, was there any external check on the validity of the scenarios? What happened if a role-player accidentally under-performed or over-performed on a scenario?

In our limitations on page 17 line 303 -306 we note the challenge incurred with local role play and suggest having video capture or formally scripted scenarios to ensure consistency and validity of the performance level.

"Our study was limited by lack of formal training and experience in role-playing by simulated learners. Our 'actors' were not professionally trained (but rather research clinicians!) and scenarios and levels

were de novo; ideally, with more resources and time, mock scenarios would be formally scripted and/or video-captured to optimize standardization."

Throughout the paper, the use of the terms "average" "acceptable" and "passing" are used interchangeably. This causes some difficulty for the reader, and standardizing this would help. I wonder if a better set of terms would be "not proficient" "proficient" and "advanced proficiency" or something like this. Clarity on this point would help Tables 2-3 especially be easier to read.

Although we agree with this perspective and actually discussed doing this, it was a team decision to use course language in describing the results. These are the terms used in the course and other like publications with HBB, ECEB and BAB.

Missing in the results is a discussion of how the "proficient" scenarios were misclassified - were they under- or over-rated? This seems like an important point

We appreciate this comment and have addressed it by adding in the results, page 12 line 211:

Raters were more accurate in identifying 'poor' and 'excellent' compared to average, *and often identified excellent proficiency level scenarios as average.*

Under "what is known" - don't use abbreviations here either. All thre of the "what this study adds" statements are just statements about what is not known. I would replace at least two with positive statements of what is known. It's ok to point out that gaps in knowledge exist, but, for example, there are some studies that measure IRR and we already know that a moderate IRR can be obtained.

A fair comment and we have addressed this by replacing two of the statements in this section: (page 19)

1. ***Studies examining the effectiveness of Helping Babies Breathe, Essential Care for Every Baby and Bleeding after Birth report improvements in clinician skills post training.***
2. ***Global partners support course evaluations in most published studies.***
3. ***Experts in the field recommend that all examiners undergo rater training prior to becoming an OSCE assessor***.

   Under "What this study adds" - The point about misclassification of proficient/average performance is the most important finding and should be mentioned.

We added an extra statement to address this, moving the original statement #3 to #4. (Page 19)

The new #3 reads: Raters often identified excellent proficiency as average.

1. A conceptual framework for training in country health providers as raters in an LMIC
2. ***Raters had a hard time identifying average performance, despite the achievement of moderate rater agreement.***
3. Raters often identified excellent proficiency as average.
4. OSCE checklist multi-step items add complexity and should be adapted to a local context