RATER TRAINING FOR STANDARDISED ASSESSMENT OF OBJECTIVE STRUCTURED CLINICAL EXAMINATIONS IN RURAL TANZANIA

Sigalet, Elaine L;Matovelo, Dismas;Brenner, Jennifer L;Boniphace, Maendeleo;Ndaboine, Edgar;Mwaikasu, Lusako;Shabani, Girles;Mannerfeldt, Jaelene;Singhal, Nalini;

;

IDRC Grant: 108024-001-Replicating the MamaToto Program in Rural Tanzania (IMCHA)

BMJ Paediatrics Open

Rater training for standardised assessment of Objective Structured Clinical Examinations in rural Tanzania

Elaine L Sigalet ,[1] Dismas Matovelo,[2] Jennifer L Brenner,[3] Maendeleo Boniphace,[2] Edgar Ndaboine,[2] Lusako Mwaikasu,[2] Girles Shabani,[2] Julieth Kabirigi,[2] Jaelene Mannerfeldt,[1] Nalini Singhal[1]

To cite: Sigalet EL, Matovelo D, Brenner JL, *et al*. Rater training for standardised assessment of Objective Structured Clinical Examinations in rural Tanzania. *BMJ Paediatrics Open* 2020;**4**:e000856. doi:10.1136/bmjpo-2020-000856

Received 31 August 2020
Revised 16 November 2020
Accepted 20 November 2020

[1]Community Health Sciences, University of Calgary Cumming School of Medicine, Calgary, Alberta, Canada
[2]Obstetrics and Gynecology, Nursing, Pediatrics, Catholic University of Health and Allied Sciences, Bugando Medical Center, Mwanza, Mwanza, Tanzania
[3]Faculty of Medicine, University of Calgary Cumming School of Medicine, Calgary, Alberta, Canada

Correspondence to
Dr Elaine L Sigalet; elaine.sigalet@gmail.com

ABSTRACT

Objectives To describe a simulation-based rater training curriculum for Objective Structured Clinical Examinations (OSCEs) for clinician-based training for frontline staff caring for mothers and babies in rural Tanzania.

Background Rater training for OSCE evaluation is widely embraced in high-income countries but not well described in low-income and middle-income countries. Helping Babies Breathe, Essential Care for Every Baby and Bleeding after Birth are standardised training programmes that encourage OSCE evaluations. Studies examining the reliability of assessments are rare.

Methods Training of raters occurred over 3 days. Raters scored selected OSCEs role-played using standardised learners and low-fidelity mannikins, assigning proficiency levels a priori. 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.

Results Six healthcare providers scored 42 training scenarios. There was moderate rater agreement across all OSCEs (κ=0.508). Kappa values increased with Helping Babies Breathe (κ=0.28–0.48) and Essential Care for Every Baby (κ=0.42–0.77) by day 3 of training, but not with Bleeding after Birth (κ=0.58–0.33). Raters identified average proficiency 50% of the time.

Conclusion Our study shows that the in-country raters in this study had a hard time identifying average performance despite moderate rater agreement. Rater training is critical to ensure that the potential of training programmes translates to improved outcomes for mothers and babies; more research into the concepts and training for discernment of competence in this setting is necessary.

What is known about the subject?

- Studies examining the effectiveness of Helping Babies Breathe, Essential Care for Every Baby and Bleeding after Birth report improvements in clinician skill post-training.
- Global partners support course evaluations in most published studies.
- Experts in the field recommend that all examiners undergo rater training prior to becoming an Objective Structured Clinical Examination (OSCE) assessor.

What this study adds?

- A conceptual framework for training in-country health providers as raters in a low/middle-income country.
- Raters had a hard time identifying average performance, despite the achievement of moderate rater agreement.
- Raters often identified excellent proficiency as average.

BACKGROUND

Helping Babies Breathe (HBB) and Essential Care for Every Baby (ECEB), from the Helping Babies Survive (HBS) programme[1 2] and the Bleeding after Birth (BAB) from the Helping Mothers Survive (HMS) programme[3 4] are examples of standardised health provider training programmes designed by expert clinicians and educators from high-income countries (HICs) with input from low/middle-income countries (LMICs) for use in LMICs. The HBB training course reviews skills related to newborn resuscitation; ECEB focuses on newborn routine care and danger sign identification; BAB reviews management of maternal haemorrhage. All three courses and others in the HMS, HBS series use low-fidelity manikins, hands-on simulation practice of common case scenarios and emphasise compliance with algorithm-based 'Action Plans'. Course content addresses common gaps that lead to some of the highest sources of global maternal[5 6] and newborn mortality.[1 2]

The competence of participants in these courses is frequently assessed using Objective Structured Clinical Examinations (OSCEs). A number of studies in a variety of LMIC settings have demonstrated improvements in provider competency managing relevant obstetric and neonatal cases post-training.[6–16] However, few of these studies provide details of assessor training, or the reliability of the

OSCE assessments.[10 15 16] Furthermore, only one study used in-country OSCE raters[15]; others have relied on external (from outside the country of study) development and academic partners serving in-rater roles.[9 16] Training of raters to serve as OSCE assessors is widely embraced in HIC,[17–25] but rater training has not been well described in LMICs. Reisman *et al* refer to standardised OSCE training but do not report details.[15] Formal pre-OSCE training for assessors aims to minimise sources of measurement error,[17–25] increasing confidence that a participant's OSCE score truly reflects their competence. With OSCE administration, sources of error can arise from the OSCE structure and/or rater objectivity.[17 19 22 25] Facilitator materials for HBB, ECEB and BAB courses provide clear guidelines to minimise measurement error with the OSCE administration. For example, Jhpeigo provides information on quality assessment[3] for their HMS training series, but there are no guidelines for training OSCE raters or evaluating rater agreement. The purpose of our study was to describe a simulation-based OSCE rater training curriculum and assessment of subsequent levels of rater agreement with administration of OSCEs in rural Tanzania using locally trained healthcare providers as raters.

METHOD

Patient and public involvement

Patients were not involved in this study.

Setting

The study was conducted in Kwimba District located in Mwanza Region, Tanzania over 3 days; 2 days in April 2018 and 1 day in May 2018.

Participants

Raters were recruited from clinical staff practising in the rural health facilities in the district where training was to occur. 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. All selected participants provided informed consent to be involved in the study. Rater characteristics and rater OSCE scores for each OSCE scenario were collated under a master tracking number to ensure rater anonymity. Following 3 days of rater training, participants were involved as raters for OSCE evaluations to assess workshop learners pretraining and post-training, at 6 and at 12 months. The rater training curriculum was led by a team comprised of clinician researchers from Catholic University of Health and Allied Sciences and University of Calgary.

Design

This study used a descriptive study design (figure 1). Raters attended rater training prior to any formal scoring of workshop participants. Categorical levels of proficiency (poor, acceptable and excellent) (decided a priori) were role-modelled by clinician research team members for each OSCE each day to create a mock scoring context. All six raters observed and scored the exact same scenario at the same time, making judgements about observed behaviours independent of discussion with each other. Scores were collected and then reviewed with the raters; areas of disagreement were explored, using an inquiry approach for debriefing. Zabar's review criteria and mitigation strategies were used as the framework for both the reviews and refining methodology. The research team lead (content expert) gave direct feedback. Categorical levels of proficiency that challenged rater agreement

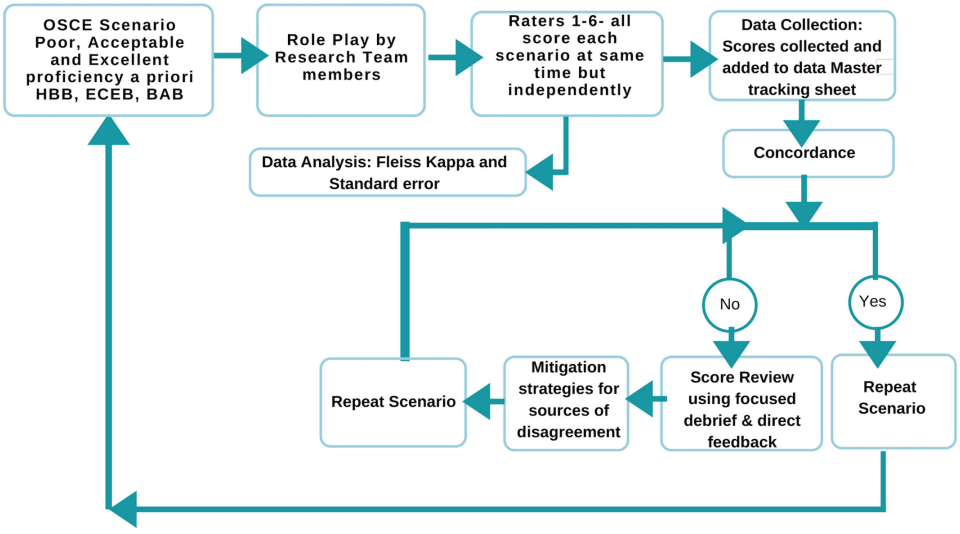

Figure 1 This figure provides a visual of the research design we used in the study each day. All six raters scored all 42 of the role-played scenarios with proficiency determined a priori. Raters participated in 42 debrief sessions over the 3 days. BAB, Bleeding after Birth; ECEB, Essential Care for Every Baby; HBB, Helping Babies Breathe; OSCE, Objective Structured Clinical Examination.

Table 1 Proficiency level identification (number of scenarios and resulting percentage correctly identified in proficiency category)

OSCE	Proficiency level	n	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6
All		42						
	Poor	10	10 (100%)	10 (100%)	10 (100%)	10 (100%)	10 (100%)	9 (90%)
	Average	18	8 (44%)	9 (50%)	9 (50%)	5 (28%)	8 (44%)	8 (44%)
	Excellent	14	10 (71%)	7 (50%)	10 (71%)	8 (57%)	11 (79%)	10 (71%)
BAB		15						
	Poor	6	6 (100%)	6 (100%)	6 (100%)	6 (100%)	6 (100%)	5 (83%)
	Average	3	2 (66%)	0 (0%)	0 (0%)	0 (0%)	2 (66%)	2 (66%)
	Excellent	6	3 (50%)	1 (17%)	3 (50%)	2 (33%)	5 (83%)	3 (50%)
ECEB		12						
	Poor	2	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)
	Average	6	0 (0%)	3 (50%)	3 (50%)	1 (17%)	2 (33%)	3 (50%)
	Excellent	4	4 (100%)	4 (100%)	4 (100%)	4 (100%)	4 (100%)	4 (100%)
HBB		15						
	Poor	2	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)
	Average	9	6 (67%)	6 (67%)	6 (67%)	4 (44%)	4 (44%)	3 (33%)
	Excellent	4	3 (75%)	2 (50%)	3 (75%)	2 (50%)	2 (50%)	3 (75%)

BAB, Bleeding after Birth; ECEB, Essential Care for Every Baby; HBB, Helping Babies Breathe ; OSCE, Objective Structured Clinical Examination.

were repeated. Checklists were collected and collated on an MS Excel spreadsheet on a research-dedicated computer. Field notes were used to track challenges. SPSS V.26 was used to analyse rater data. Descriptive statistics were used to provide information about mock scoring and rater's abilities to identify the three categorical levels of proficiency. All raw scores indicating excellent levels of proficiency (table 1) were also analysed as acceptable (table 2) to align with training programme guidelines: two categories of proficiency. Fleiss' kappa with SE was calculated to provide information about the level of rater agreement.[26] Kappa values of <0.20, 0.21–0.40, 0.41–0.60, 0.61–0.80 and 0.81–1.00 are considered poor, fair, moderate, good and very good, respectively.[26]

Evaluation tools

The OSCEs used were drawn from training programme materials.[1–4] There were 24 pass/fail items on the HBB

Table 2 Proficiency level identification: (average and excellent categories combined) for training programme categories (number of scenarios and resulting percentage correctly identified in proficiency category)

OSCE	Proficiency level	n	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6
All		42						
	Poor	10	10 (100%)	10 (100%)	10 (100%)	10 (100%)	10 (100%)	9 (90%)
	Average	32	18 (56%)	16 (50%)	19 (59%)	14 (44%)	19 (59%)	18 (56%)
BAB		15						
	Poor	6	6 (100%)	6 (100%)	6 (100%)	6 (100%)	6 (100%)	5 (83%)
	Average	9	5 (55%)	1 (11%)	3 (33%)	2 (22%)	7 (78%)	5 (66%)
ECEB		12						
	Poor	2	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)
	Average	10	4 (40%)	7 (70%)	7 (70%)	5 (50%)	6 (60%)	3 (50%)
HBB		15						
	Poor	2	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)
	Average	13	9 (69%)	8 (62%)	9 (69%)	6 (46%)	6 (46%)	6 (46%)

BAB, Bleeding after Birth; ECEB, Essential Care for Every Baby; HBB, Helping Babies Breathe ; OSCE, Objective Structured Clinical Examination.

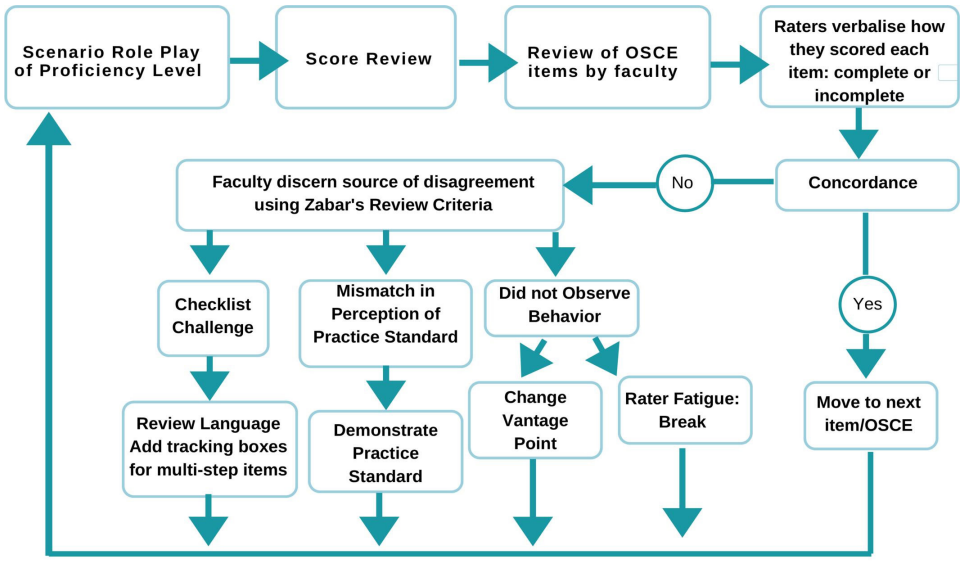

Figure 2. This figure provides a visual of the Conceptual framework used to improve the level of rater agreement.

Figure 2 This figure provides a visual of the conceptual framework used to improve the level of rater agreement. OSCE, Objective Structured Clinical Examination.

OSCE, 15 items on the ECEB OSCE and 14 items on the BAB OSCE. All raters were familiar with the OSCE checklists and relevant training course content as they had recently participated in the same courses themselves as learners. Poor proficiency, often referred to as 'red' in reported studies was identified by a score of <71%: 0–17, 0–10 and 0–9 on the HBB, ECEB and BAB OSCE, respectively. Learner scores >70% identified an 'acceptable' level of proficiency or 'green' in reported studies: >17, >10 and >9 on HBB, ECEB and BAB, respectively.[1–4] The research team added a third category, a candidate's score of >22, >13 and >12 identified excellent proficiency for HBB, ECEB and BAB OSCEs, respectively. To standardise 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.

The rater curriculum

The conceptual framework (figure 2) and Zabar's review criteria[27] provide details about elements of the curriculum and the iterative nature of the training process. Three physical OSCE stations were set up to facilitate learner transition between each testing station. 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 learnt 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.

RESULTS

Raters (n=6) included physicians (n=1), midwives (n=4) and nurses (n=1). All study participants completed the 3 full days of rater training which included participation in scoring and a focused debrief for 42 scenarios over the 3 days. Table 3 provides details about scenario scoring for HBB, ECEB and AMSTL over the 3 days.

The time needed for each OSCE station with score review was longer for average proficiency levels (30–40 min) when compared with 'excellent' and 'poor' proficiency levels (15–20 min). Fleiss' kappa values (table 3) showed that there was a moderate level of rater agreement in identifying 'poor' and 'acceptable' proficiency across all OSCEs (κ=0.51). Kappa values improved over the 3 days moving from 'fair' to 'moderate' for the HBB OSCE and 'moderate' to 'good' for the ECEB OSCE. The kappa value for BAB was 'moderate' on day 1 but decreased to 'fair' on day 2 and day 3. Information about rater abilities to correctly identify proficiency levels is described in tables 1 and 2.

Raters were more accurate in identifying 'poor' and 'excellent' compared with average, and often identified excellent proficiency level scenarios as average. Raters identified average proficiency approximately 50% of the time (tables 1 and 2). Information detailing challenges from field notes is presented in table 4.

DISCUSSION

This study describes an OSCE rater training curriculum and presents evaluation of the curriculum showing levels of rater agreement for HBB, ECEB and BAB training courses in an LMIC. Quality rater training and subsequent reliability analysis are especially important in LMIC context because of the limited quality assurance

Table 3 Kappa values

Training programme	Proficiency level	n	Average			Day 1 (n=16)			Day 2 (n=14)			Day 3 (n=12)	
			Fleiss' κ	SE	n	Fleiss' κ	SE	n	Fleiss' κ	SE	n	Fleiss' κ	SE
HBB		15	0.43	0.07		0.28	0.12		0.58	0.12		0.48	0.12
	Poor	2			0			1			1		
	Acceptable	13			5			4			4		
ECEB		12	0.61	0.07		0.42	0.10		0.70	0.13		0.77	0.15
	Poor	2			1			1			0		
	Acceptable	10			4			3			3		
BAB		15	0.46	0.07		0.58	0.12		0.19	0.12		0.33	0.12
	Poor	6			2			2			2		
	Acceptable	9			3			3			3		
All OSCEs			0.508	0.04									

BAB, Bleeding after Birth; ECEB, Essential Care for Every Baby; HBB, Helping Babies Breathe; OSCEs, Objective Structured Clinical Examinations.

monitoring patient safety in the system and resources.[28–30] Our results suggest that the moderate levels of rater agreement, coupled by notable challenges in discriminating 'acceptable' performance, expose a potential for either overestimating or underestimating competence. This has consequences for the individual, the training programme and the system. The challenge incurred in discriminating between borderline performance is not isolated to an LMIC context but reported universally.[31–33] With overestimation of competence, training programmes 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 utilisation: with underestimation of competence, training programmes may be directing the limited resources to clinicians who do not need extra training. Further, frontline staff frequently work short staffed when someone is away at training, so that unnecessary remediation training may exacerbate staff overload.[28–30 34]

In the majority of HBB, ECEB and BAB training programme reports, validation of improved caregiver competency is determined by comparing pretraining and post-training OSCE scores. Our results suggest that the existing reports describing a moderate inter-rater reliability (IRR) may be misleading without further validation of the accuracy of rater discernment of acceptable proficiency.[10 15 16] Our raters achieved moderate rater agreement yet discernment of acceptable proficiency, which is the pass criterion in these training programmes, was approximately 50%. Based on our findings we would suggest including both measures of validation. Considering contexts with limited resources, it may be helpful to implement a further strategy such a global rating scale, which is common practice in HICs[17–19 22–25] to provide another method of validation of participant competence.[26 27] A Global Rating Scale allows the rater to evaluate how well a learner performs on a scale of 1–5, with 5 reflecting the highest level of competence.[27] More than one method of validation creates more certainty that results are an accurate reflection of participant competence and/or training programme efficacy.[26] With the continued high reports of maternal and neonatal mortality, it is important to be confident that these training programmes are accurate in identifying and supporting clinicians who may not be providing safe care on the frontline.

Table 4 Rater challenges from field notes

Challenge	HBB	ECEB	BAB
Differing perceptions of practice standard		Back rub stimulation Sequence for drying baby	Fundal massage Bleeding assessment Frequency of bleeding assessment
Tracking multistep OSCE items	Item 1. Prepares area for delivery	Item 7. Improves thermal care	Item 7. Controlled cord traction counter pressure
	Item 2. Equipment preparation Item 3. Hand washing Item 5. Removes wet clothes Item 24. Communication and teaching	Item 8. Identifying danger signs Advanced care classification Item 10. Medication calculation and administration	Item 12. Determining postpartum haemorrhage
OSCE English words		Hypothermia	Hypertension
Actions without verbalising		Warming baby	

BAB, Bleeding after Birth; ECEB, Essential Care for Every Baby; HBB, Helping Babies Breathe; OSCE, Objective Structured Clinical Examination.

The guidelines for OSCE rater training used in this study were based on recommendations from HIC rater training experiences; these are challenging to implement in an LMIC context. Globally, good practice is for OSCE raters to have relevant content expertise, be well orientated to the OSCE checklist and use a validated rating scale.[22–25] Although we strove for this, we had a limited pool of potential raters; this may have affected the challenges we noted in rater perceptions of the expected practice standard. Raters were recruited by clinician researchers based on recollections of which previous participants from recent HBB, ECEB and BAB trainings had performed well; no objective strategy was employed in their selection. This was the reason in-country faculty inserted a third categorical level of proficiency: excellent. They wanted an objective strategy to identify content experts as the future raters for such training programmes. A quality rater training curriculum includes standardised mock scenarios where raters practise with a variety of expected learner proficiency levels demonstrated and practice scored. In our study, this was one of the greatest challenges. Research clinicians' role-playing scenarios on day 1 were challenged in demonstrating poor proficiency. In discussion, they shared they did not want participants to think they were not experts in the field. The inclusion of scripted and video capture of proficiency levels may lessen this tension and inconsistency in role-play. Despite this, the level of rater agreement improved over the 3 training days for both HBB and ECEB. The fall-off in rater agreement for BAB on day 3 was unexpected but may be in part related to the timing of these scenarios on day 3; they were the last role-plays of the day and rater fatigue may have played a role. Additionally, the greater number of differing perceptions of the practice standard (table 4) may have impacted this finding.

A solid rater curriculum incorporates a framework such as Zabar's (figure 2) to guide rater feedback; this is especially important in a setting where the concept of rater training is novel. In our study, Zabar's framework was simple and easy to use as evidenced by a decreased level of external coaching each day. A study strength was the achievement of a level of rater agreement similar to the few published training course reports for ECEB and HBB. In our participant group, the 'moderate to good' kappa for the ECEB OSCE was as reported by Kassick *et al* in Ghana, the only other ECEB reported study to include in-country evaluators: a regional and national evaluator.[10] In the HBB OSCE, our findings demonstrated 'fair to moderate' kappa value which was similar to the 'fair to good' kappa value reported by Reisman *et al* in Tanzania[15] whose raters included two external evaluators and one country-based evaluator. Comparable studies for kappa value results for raters scoring the BAB OSCE module are not reported. The achievement of comparable IRR to the studies using in-country and external partners provides support for the rater training curriculum, yet the inability to accurately discern acceptable proficiency (pass criteria) is concerning. To gain further insight into the relationship between faculty role-play and the inability to discern acceptable proficiency, we plan to script the acceptable proficiency level for each OSCE, coach faculty in the role-play, and repeat the curriculum and analysis.

Rater trainees were challenged by OSCE items where scores incorporated multisteps for their achievement; this was consistent with experiences described by Seto *et al* who also identified lower rater agreement for HBB OSCE multistep items.[16] For example, in our study, one HBB OSCE 'item' requires the learner to 'prepare the area for delivery'. To achieve a point and 'pass' this item, the learner must complete all four of the following: (1) place towels at bedside; (2) place suction at bedside; (3) place a bag and mask at bedside; and (4) place oxytocin at bedside. This 'item' created confusion among rater trainees; during mock session review, several participants had 'passed' the mock scenario learner on this item despite not having seen all steps yet having observed at least one step. To address this gap, we added subitem tracking boxes when this challenge was identified on day 1; the use of this strategy warrants further study.

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 optimise standardisation. Additionally, time constraints necessitated working 3 long days; rater fatigue was likely. This was especially true for one pregnant rater-trainee who participated for the first 2 days then arrived with newborn in hand on day 3. Our results may have limitations in generalisability but do provide some context and learning for others interested in developing a rater training curriculum in a low-resource setting.

CONCLUSION

Our results 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. Clinician everywhere need ongoing training, but to optimise 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 standardisation 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. Global training programmes, including HBB, ECEB and the BAB need to be confident that OSCE scores truly reflect learner ability, to identify and support those needing further skill practice. Significant global investments have been made towards

maternal newborn health provider training; participants need to leave workshop venues equipped with the skills to save mothers' and newborns' lives. We hope this experience encourages programme developers nationally and internationally to scale up in-country rater training. For LMIC simulation-based training programmes to be sustainable, all countries and regions should have their own trained OSCE raters.

Acknowledgements Healthcare workers from Misungwi District who served as raters for this training programme.

Contributors ELS, DM, JLB and NS provided substantial contributions to the conception and design of the work, drafting and revising the manuscript, approved the submitted version and agree to be accountable for aspects of the work related to accuracy or integrity of any part of the work. GS contributed substantially to acquisition and analysis of data, and revision of manuscript drafts, approved the submitted version and agree to be accountable for all aspects of the work ensuring questions related to accuracy or integrity are examined and resolved. MB, EN, LM and JK contributed substantially to interpretation of data and revision of manuscript drafts, approved submitted version and agree to be accountable for all aspects of the work ensuring questions related to accuracy or integrity are examined and resolved. JM contributed substantially to conception of work and revision of submitted manuscript, approved submitted manuscript and agrees to be accountable ensuring questions related to accuracy or integrity are examined and resolved.

Funding This work was supported by a grant from the Innovating for Maternal and Child Health in Africa (IMCHA) initiative, a partnership of Global Affairs Canada (GAC), the Canadian Institutes of Health Research (CIHR) and Canada's International Development Research Centre (IDRC), grant number 108 024-001 under Mama na MToto programme in rural Tanzania.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not required.

Ethics approval This study was embedded within a Simulation Enhanced Maternal Newborn Health training workshop, conducted as part of an ongoing rural education programme. The study was approved by Catholic University of Health and Allied Sciences Ethics Board (#CREC/070/2015), the Tanzania National Institute for Medical Research (NIMR) (#MR/53/100/525), and University of Calgary Science and Ethics Board (#REB15-1919).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplemental information. Deidentified participant data are presented in tables. Database is secured by the University of Calgary which is monitored by the Global Health Research Office; hf.mercader@ucalgary.ca.

ORCID iD

Elaine L Sigalet http://orcid.org/0000-0002-9103-9968

REFERENCES

1 American Academy of Paediatrics. *Guide for Implementation of Helping Babies Breathe(HBB): Strengthening neonatal resuscitation in suitable programs of essential newborn care*, 2011.
2 American Academy of Pediatrics. Helping babies breathe, 2nd edition, 2015. Available: https://www.aap.org/en-us/advocacy-and-policy/aap-health-initiatives/helping-babies-survive/Pages/Helping-Babies-Breathe.aspx [Accessed 19 Mar 2018].
3 Bluestone J, Fowler R, Johnson P, *et al*. *Jhpeigo helping mothers survive training skills for health care providers, third edition: reference manual*. Jhpeigo Corporation, 2010. http://resources.jhpiego.org/system/files/resources/trainingskills_manual_0.pdf
4 Jhpeigo. Helping mothers survive bleeding after birth training package, 2016. Available: http://reprolineplus.org/resources/HMS [Accessed 19 Mar 2018].
5 Department of Reproductive Health and research,. *World Health organization (who) who recommendations for the prevention and treatment of postpartum hemorrhage*. Geneva: WHO, 2012.
6 Evans CL, Johnson P, Bazant E, *et al*. Competency-based training "Helping Mothers Survive: Bleeding after Birth" for providers from central and remote facilities in three countries. *Int J Gynaecol Obstet* 2014;126:286–90.
7 Kamath-Rayne BD, Thukral A, Visick MK, *et al*. Helping babies breathe, second edition: a model for strengthening educational programs to increase global newborn survival. *Glob Health Sci Pract* 2018;6:538–51.
8 Nelissen E, Ersdal H, Ostergaard D, *et al*. Helping mothers survive bleeding after birth: an evaluation of simulation-based training in a low-resource setting. *Acta Obstet Gynecol Scand* 2014;93:287–95.
9 Brucker MC. Management of the third stage of labor: an evidence-based approach. *J Midwifery Womens Health* 2001;46:381–92.
10 Kassick ME, Chinbuah MA, Serpa M, *et al*. Evaluating a novel neonatal-care assessment tool among trained delivery attendants in a resource-limited setting. *Int J Gynaecol Obstet* 2016;135:285–9.
11 Alwy Al-Beity F, Pembe A, Hirose A, *et al*. Effect of the competency-based *Helping Mothers Survive Bleeding after Birth* (HMS BAB) training on maternal morbidity: a cluster-randomised trial in 20 districts in Tanzania. *BMJ Glob Health* 2019;4:e001214.
12 Bishanga DR, Charles J, Tibaijuka G, *et al*. Improvement in the active management of the third stage of labor for the prevention of postpartum hemorrhage in Tanzania: a cross-sectional study. *BMC Pregnancy Childbirth* 2018;18:233.
13 Ameh CA, van den Broek N. Making it happen: training health-care providers in emergency obstetric and newborn care. *Best Pract Res Clin Obstet Gynaecol* 2015;29:1077–91.
14 Niermeyer S. From the neonatal resuscitation program to helping babies breathe: global impact of educational programs in neonatal resuscitation. *Semin Fetal Neonatal Med* 2015;20:300–8.
15 Reisman J, Martineau N, Kairuki A, *et al*. Validation of a novel tool for assessing newborn resuscitation skills among birth attendants trained by the helping babies breathe program. *Int J Gynaecol Obstet* 2015;131:196–200.
16 Seto TL, Tabangin ME, Josyula S, *et al*. Educational outcomes of helping babies breathe training at a community hospital in Honduras. *Perspect Med Educ* 2015;4:225–32.
17 Khan KZ, Ramachandran S, Gaunt K, *et al*. The objective structured clinical examination (OSCE): AMEE guide No. 81. Part I: an historical and theoretical perspective. *Med Teach* 2013;35:e1437–46.
18 Roberts C, Newble D, Jolly B, *et al*. Assuring the quality of high-stakes undergraduate assessments of clinical competence. *Med Teach* 2006;28:535–43.
19 Humphrey-Murto S, Touchie C, Smee S. Oxford Textbook of Medical Education. Chapter 45. In: *Objective structured clinical examinations*. Oxford UK: Oxford University Press, 2013.
20 van der Vleuten CP, van Luyk SJ, van Ballegooijen AM, *et al*. Training and experience of examiners. *Med Educ* 1989;23:290–6.
21 Harden RM. Revisiting 'Assessment of clinical competence using an objective structured clinical examination (OSCE)'. *Med Educ* 2016;50:376–9.
22 Feldman M, Lazzara EH, Vanderbilt AA, *et al*. Rater training to support high-stakes simulation-based assessments. *J Contin Educ Health Prof* 2012;32:279–86.
23 Schleicher I, Leitner K, Juenger J, *et al*. Examiner effect on the objective structured clinical exam – a study at five medical schools. *BMC Med Educ* 2017;17:1–7.
24 Daniels VJ, Pugh D. Twelve tips for developing an OSCE that measures what you want. *Med Teach* 2018;40:1208–13.
25 Preusche I, Schmidts M, Wagner-Menghin M. Twelve tips for designing and implementing a structured rater training in OSCEs. *Med Teach* 2012;34:368–72.
26 Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas* 1973;33:613–9.
27 Zabar S, Krajic Kacher E, Kalet A, *et al*. *Objective structured clinical exams- 10 steps to planning and implementing OSCE's and other standardized patient exercises*. New York: Springer, 2013.
28 World Health Organization, OECD, International Bank for Reconstruction and Development/The World Bank. *Delivering quality health services. A global imperative for universal health coverage*, 2018.

29 Kruk ME, Gage AD, Arsenault C, *et al*. High-quality health systems in the sustainable development goals era: time for a revolution. Available: https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(18)30386-3/fulltext [Accessed 6 Nov 2018].
30 Rowe AK, Labadie G, Jackson D, *et al*. Improving health worker performance: an ongoing challenge for meeting the sustainable development goals. *BMJ* 2018;362:k2813.
31 Fuller R, Homer M, Pell G, *et al*. Managing extremes of assessor judgment within the OSCE. *Med Teach* 2017;39:58–66.
32 Petrusa ER. Clinical Performance Assessments. In: *International handbook of reserach in medical education Boston*. Kluwer Academic Publishers, 2002: 673–709.
33 Reid K, Smallwood D, Collins M, *et al*. Taking OSCE examiner training on the road: reaching the masses. *Med Educ Online* 2016;21:32389.
34 The United Republic of Tanzania Ministry of Health and Social Welfare. Health sector strategic plan July 2015-June 2020: reaching all households with quality care. Available: https://dc.sourceafrica.net/documents/118198-Tanzania-Health-Sector-Strategic-Plan-July-2015.html [Accessed 19 Mar 2017].