

# BMJ Paediatrics Open

BMJ Paediatrics Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Paediatrics Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjpaedsopen.bmj.com>).

If you have any questions on BMJ Paediatrics Open's open peer review process please email [info.bmjpo@bmj.com](mailto:info.bmjpo@bmj.com)

# BMJ Paediatrics Open

## Rater training for standardized assessment of Objective Structured Clinical Exams in rural Tanzania

Journal:	<i>BMJ Paediatrics Open</i>
Manuscript ID	bmjpo-2020-000856
Article Type:	Original research
Date Submitted by the Author:	31-Aug-2020
Complete List of Authors:	Sigalet, Elaine; University of Calgary Cumming School of Medicine, Community Health Sciences Matovelo, Dismas; Catholic University of Health and Allied Sciences Boniphace, Maendeleo; Catholic University of Health and Allied Sciences Shabani, Girles; Catholic University of Health and Allied Sciences Ndaboine, Edgar; Catholic University of Health and Allied Sciences Mwaikasu, Lusako; Catholic University of Health and Allied Sciences Kabiligi, Julieth; Catholic University of Health and Allied Sciences Brenner, Jennifer; University of Calgary Cumming School of Medicine, Faculty of Medicine Mannerfeldt, Jaelene; University of Calgary Cumming School of Medicine, Community Health Sciences Singhal, Nalini; University of Calgary Cumming School of Medicine, Community Health Sciences
Keywords:	Health services research, Resuscitation

SCHOLARONE™  
Manuscripts



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

1  
2  
3 Rater training for standardized assessment of Objective Structured Clinical Exams in rural  
4  
5 Tanzania  
6

7  
8 **Authors:** Elaine Sigalet<sup>1</sup>, Dismas Matovelo<sup>2</sup>, Maendeleo Boniphace<sup>2</sup>, Girles Shabani<sup>2</sup>, Edgar  
9  
10 Ndaboine<sup>2</sup>, Lusako Mwaikasu<sup>2</sup>, Julieth Kabiligi<sup>2</sup>, Jennifer L Brenner<sup>1</sup>, Jaelene Mannerfeldt<sup>1</sup>,  
11  
12 Nalini Singhal<sup>1</sup>  
13

14  
15 **Institution Affiliations:**

16  
17 <sup>1</sup>University of Calgary, Cummings School of Medicine, Alberta  
18  
19 Canada  
20

21 <sup>2</sup>Catholic University of Health & Allied Sciences, Tanzania  
22

23  
24 **Corresponding author:** Elaine Sigalet [E-mail: [elaine.sigalet@gmail.com](mailto:elaine.sigalet@gmail.com)], Department of  
25  
26 Community Health Sciences, University of Calgary Cummings School of Medicine, 3330 Hospital  
27  
28 Drive NW. Calgary, AB, Canada T2N4N1, Tel.+587-4386604  
29

30  
31 **Co-Authors:**

32  
33 **Dismas Matovelo**, Department of Obstetrics and Gynecology, Catholic University of Health &  
34  
35 Allied Sciences, Bugando Medical Center, Mwanza, Tanzania  
36

37  
38 **Maendeleo Boniphace**, School of Nursing, Catholic University of Health & Allied Sciences,  
39  
40 Bugando Medical Center, Mwanza, Tanzania  
41

42  
43 **Girles Shabani**, Research Coordinator Mama na Mtoto Project, Catholic University of Health &  
44  
45 Allied Sciences, Bugando Medical Center, Mwanza, Tanzania  
46

47  
48 **Edgar Ndaboine**, Department of Obstetrics and Gynecology, Catholic University of Health &  
49  
50 Allied Sciences, Bugando Medical Center, Mwanza, Tanzania  
51

52  
53 **Lusako Mwaikasu**, Department of Obstetrics and Gynecology, Catholic University of Health &  
54  
55 Allied Sciences, Bugando Medical Center, Mwanza, Tanzania  
56  
57  
58  
59  
60

1  
2  
3 **Julieth Kabiligi**, Department of Pediatrics, Catholic University of Health & Allied Sciences,  
4 Bugando Medical Center, Mwanza, Tanzania  
5  
6

7  
8 **Jennifer L Brenner**, Department of Pediatrics and Community Health Sciences, Director Global  
9 Maternal Newborn Child Health, University of Calgary Cummings School of Medicine, Calgary,  
10 Alberta Canada  
11  
12

13  
14 **Jaelene Mannerfeldt**, Department of Obstetrics and Gynecology, , University of Calgary  
15 Cummings School of Medicine, Calgary, Alberta Canada  
16  
17

18  
19 **Nalini Singhal**, Department of Neonatology, University of Calgary Cummings School of  
20 Medicine, Calgary, Alberta Canada  
21  
22  
23  
24  
25

#### 26 **Contributor Statements:**

27  
28 Elaine Sigalet, Dismas Matovelo, Jennifer L Brenner and Nalini Singhal provided substantial  
29 contributions to the conception and design of the work, drafting and revising the manuscript,  
30 approve the submitted version and agree to be accountable for aspects of the work related to  
31 accuracy or integrity of any part of the work.  
32  
33

34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Girls Shabani contributed substantially to acquisition and analysis of data, revision of manuscript drafts, approve submitted version and agree to be accountable for all aspects of the work ensuring ensuring questions related to accuracy or integrity are examined and resolved.

Maendeleo Boniphace, Edgar Ndaboine, Lusako Mwaikasu, and Julieth Kabiligi contributed substantially to interpretation of data, revision of manuscript drafts, approve submitted version and agree to be accountable for all aspects of the work ensuring questions related to accuracy or integrity are examined and resolved.

1  
2  
3 Jaelene Mannerfeldt contributed substantially to conception of work, revision of submitted  
4 manuscript, approves submitted manuscript and agrees to be accountable ensuring questions  
5 related to accuracy or integrity are examined and resolved.  
6  
7  
8  
9

10  
11 **Key Words** Simulation Training, Community Child Health, Resuscitation, Mortality, Medical  
12 Education  
13

14  
15  
16 **Word Count: 2928**  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**Abstract:****OBJECTIVES**

To describe a simulation rater training curriculum for Objective Structured Clinical Exams (OSCEs) in Tanzania.

**BACKGROUND**

Rater training for OSCE evaluation is widely embraced in high income countries (HIC) but not well described in low and middle-income countries (LMICs). Helping Babies Breathe (HBB), Essential Care for Every Baby (ECEB) and Bleeding after Birth (BAB) are standardized training programs that encourage OSCEs evaluations. Reports of the reliability of these assessments is rare, making score inferences vulnerable.

**METHODS**

Training using these programs was conducted over three days. Healthcare providers scored selected OSCEs role played using standardized learners and low fidelity mannikins; proficiency levels were determined *a priori*. Zabar's review criteria guided rater feedback in score review. Descriptive statistics and Fleiss' kappa provided information about rater agreement. Challenges were tracked with field notes.

**RESULTS**

Six healthcare providers scored 42 training scenarios. Fleiss' kappa value shows moderate levels of rater agreement with 'poor' and 'acceptable' proficiency across all OSCEs ( $\kappa=0.508$ ,  $p<0.001$ ). Kappa values increased with HBB ( $\kappa=0.28$  to  $0.48$ ), and ECEB ( $\kappa=0.42$  to  $0.77$ ) by Day 3 of training but not with BAB ( $\kappa=0.58$  to  $0.33$ ). Raters identified average proficiency 50% of the time. OSCE items with multiple steps challenged our in-country raters.

**CONCLUSION**

1  
2  
3 Our study shows in rural, Tanzania, training of in-country raters is feasible and effective. All  
4 countries and regions should have their own trained OSCE raters. Rater training is critical to  
5 ensure that the potential of training programs translates to improved outcomes for mothers and  
6 babies.  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Confidential: For Review Only



## BACKGROUND

Helping Babies Breathe (HBB) and Essential Care for Every Baby (ECEB), from the Helping Babies Survive Program[1,2] and the Bleeding after Birth (BAB) from the Helping Mothers Survive (HMS) program[3,4] are examples of standardized health provider training programs designed by expert clinicians and educators from high income countries (HIC) with input from low and 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 hemorrhage. All three courses and others in the HMS, HBS series, use low-fidelity mannequins, hands-on simulation practice of common case scenarios and emphasize 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]

Helping Babies Breathe, ECEB and BAB workshop participants are frequently assessed using Objective Structured Clinical Exams (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 OSCE assessment reliability.[10,15,16] Furthermore, only one study used in-country OSCE raters;[15] others rely on external (from outside the country of study) development and academic partners serving in rater roles.[9,16]

Training of raters to serve as OSCEs assessors is widely embraced in HIC,[17-25] but rater training has not been well described in LMICs. Reisman and colleagues refer to standardized 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<sup>3</sup> for their HMS training series, but there are no guidelines for training OSCE raters or evaluating rater agreement. The importance of reporting on the reliability and validity of scores with OSCE administration has been well described,[17-25] with only one study providing information on rater agreement using in country assessors in an LMIC.[10] 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

This study was embedded within a Simulation Enhanced Maternal Newborn Health training workshop. 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).

### 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 two days in April 2018 and one day in May 2018.

### Participants

Raters were recruited from amongst rural health facilities in the district where training was to occur. Selection was based on their demonstrated proficiency in previous Newborn Maternal

1  
2  
3 training workshops and their experience as obstetrical health providers. All selected participants  
4 provided informed consent to be involved in the study. Rater characteristics and Rater OSCE  
5 scores for each OSCE scenario were collated under a master tracking number to ensure rater  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

training workshops and their experience as obstetrical health providers. 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 three days of rater training, participants were involved as raters for OSCE evaluations to assess workshop learners pre 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 (CUHAS) and University of Calgary.

### Design

This study used a descriptive study design (Figure 1). Scenario proficiency (poor, acceptable and excellent) was decided *a priori*, and role modelled by clinician research team members to create a mock scoring context. Each participant selected to be a rater independently scored each scenario. Raters made their own judgements of observed behaviours without consulting colleagues. Checklists were collected and collated on an MSExcel spreadsheet on a research dedicated computer. Field notes were used to track challenges. SPSS version 26 was used to analyse rater data. Descriptive statistics were used to provide information about mock scoring and raters abilities to identify the three categorical levels of proficiency. All raw scores indicating excellent levels of proficiency (Table 2) were also analyzed as acceptable (Table 3) to align with training program guidelines; two categories of proficiency. Fleiss' Kappa was calculated to provide information about the level of rater agreement.<sup>27</sup> Kappa values of <0.20, 0.21-0.40, 0.41-0.60 and 0.61-0.80 and 0.81to 1.00 are considered poor, fair, moderate, good, and very good respectively.[27]

### Evaluation Tools

The OSCEs used were drawn from training program materials.[1-4] There were 24 pass/fail items on the HBB 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, & >9 on HBB, ECEB and BAB respectively.<sup>1-4</sup> 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.

### The Rater Curriculum

The conceptual framework (Figure 2) and Zabar’s review criteria[28] provides 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.

### RESULTS

Raters ( $n=6$ ) included physicians ( $n=1$ ), midwives ( $n=4$ ) and nurses ( $n=1$ ); all study participants completed the three full days of rater training. They scored a total of 42 scenarios over the three days of training. Table one provides details about scenario scoring for HBB, ECEB and AMSTL over the three days.

**Table 1.** Kappa values with significance ( $p<0.05$ )

Training Program	Proficiency Level	n	Average		Day 1 (n=16)		Day 2 (n=14)		Day 3 (n=12)	
			Fleiss' K	p value	Fleiss' K	p value	Fleiss' K	p value	Fleiss' K	p value
HBB		15	0.43	$p<0.05$	0.28	$p<0.05$	0.58	$p<0.001$	0.48	$p<0.001$
	Poor	2	0.32	$p<0.001$						
	Acceptable	13	0.32	$p<0.001$						
ECEB		12	0.61	$p<0.05$	0.42	$p<0.001$	0.70	$p<0.001$	0.77	$p<0.001$

	Poor	2	0.63	p<0.001					
	Acceptable	10	0.63	p<0.001					
BAB		15	0.46	p<0.05	0.58	p<0.001	0.19	NS	0.33
	Poor	6	0.42	p<0.001					p<0.05
	Acceptable	9	0.38	p<0.001					
All OSCEs		42	0.508	p<0.05					

The time needed for each OSCE station with score review was longer for average proficiency levels (30-40 minutes) when compared to 'excellent' and 'poor' proficiency levels (15-20 minutes). Fleiss' Kappa values (Table 1) showed that there was a moderate level of rater agreement in identifying 'poor' and 'acceptable' proficiency across all OSCEs ( $\kappa=0.51$   $p<0.001$ ). Kappa values improved over the three 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' Day 1 but decreased to 'fair' Day 2 and Day 3. Except for the kappa value for BAB Day 2, all kappa values were statistically significant ( $p<0.05$ ). Information about rater abilities to correctly identify proficiency levels is described in Table 2 and 3.

**Table 2.** Proficiency level identification with excellent category (Number of scenarios and resulting percentage correctly identified in proficiency category)

OSCE	Proficiency Level	n	Rater 1	Rater 2	Rater 3	Rater 4	Rater5	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%)

**Table 3.** Proficiency level identification for training program categories (Number of scenarios and resulting percentage correctly identified in proficiency category)

OSCE	Proficiency Level	n	Rater 1	Rater 2	Rater 3	Rater 4	Rater5	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%)

Raters were more accurate in identifying ‘poor’ and ‘excellent’ compared to average. Raters identified average proficiency 50% of the time (Table 2 and 3). Information detailing challenges from field notes are presented in Table 4. These include differing perceptions in expected standard of practice, rater fatigue, and multi-step items.

**Table 4.** Rater Challenges from Field Notes

Challenge	HBB	ECEB	BAB
Differing perceptions of practice standard		How to stimulate baby with back rubs Sequence used to dry the baby	How to massage uterus to stop bleeding How to check for bleeding Item 14: Checks mother for bleeding for 2 hours: changed to checks mother for bleeding every 15 minutes for 2 hours
Tracking multi-step OSCE items	Item 1. Prepares area for delivery Added boxes for: towels, suction, ventilation bag and oxytocin	Item 7: Improves thermal care; Added tracking boxes for removes wet clothing, adds layer of clothing/hat, positions skin to skin, raises room temperature	Item 7: Applies counter pressure when performing controlled cord traction with a contraction: added tracking box for position of hands and one for action occurring with a contraction

<p>1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31</p>	<p>Item 2. Test equipment function; added tracking box for bag/mask and suction Item 3. Washes hands Added tracking box for HCP hands, mothers hands and others abdomen Item 5. Removes wet clothes; added tracking box for covers with clean cloth Item 24; communicates with Mother; Added tracking box for putting baby skin to skin, teaching mom to check breathing and breastfeeding</p>	<p>Item 8: Recognizes baby has a danger sign and classifies baby as needing advance care- added tracking box for states baby has danger sign and box for classifies as needing advanced care Item 10. Calculates correct dose of Ampicillin and Gentamicin; Added box for ampicillin and box for gentamycin</p>	<p>Item 12: looks/asks about amount of bleeding; Added Must do one of these: added tracking box to track looks at output in the pail, box looks at vaginal flow; box for asks mother about flow</p>
<p>32 33 34 35 36 37 38 39 40 41 42 43 44 45</p>	<p>OSCE English Words</p> <p>Actions without verbalizing</p>	<p>Hypothermia- did not interpret candidate saying mother was cold as co hypothermia</p> <p>Candidate took actions to make the baby warmer but rater marks incomplete for recognizes hypothermia</p>	<p>Hypertension- did not interpret candidate saying mother had high blood pressure as hypertension</p>

## DISCUSSION

This is the first the study to describe an OSCE rater training curriculum and present 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 is especially important in LMIC context because of the limited quality assurance monitoring patient safety in the system and

resources.[29-31] Our results suggest that the moderate levels of rater agreement, coupled by notable challenges in discriminating ‘acceptable’ versus ‘poor’ performance, exposes a potential for either overestimating or underestimating competence. Additionally, raters were challenged in discriminating ‘excellent’ versus ‘average’ performance. This has consequences for the individual, the training program, and the system. If country resources are directed to those who do not need it (overestimated) or miss those who do need it (underestimated), practising clinicians operate with less skilled health providers because some are away at training. Underestimation means that the process may have missed identifying healthcare providers who are providing unsafe care to mothers and babies. Poor competence impedes quality care.[27-29] The system uses participant scores to make decisions about training priorities, continued employment and allocation of resources, which are limited.[26,29-31] This creates further strain in an already vulnerable system.[26,29-31] The programs may need to implement a further strategy such a global health rating scale, which is common practice in the developed world,[17-19,22-25] to help define the borderline healthcare providers who need more training, and healthcare providers who have demonstrated excellent proficiency with training content to be future raters.[27,28] The challenge incurred in discriminating between borderline performance is not isolated to an LMIC context but reported universally.[32-34]

The best practices for OSCE rater training curriculum that we identified through this study reflect similar recommendations from HIC rater training experience. 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] A quality rater training curriculum includes standardized mock scenarios where raters practise with a variety of expected learner proficiency levels demonstrated and practice scored. In a study by Reid and colleagues,[34] the sole use of a satisfactory proficiency



1  
2  
3 level mock for practice limited generalizability of findings to other proficiency levels. A solid  
4 rater curriculum incorporates a framework such as Zabar's (Figure 2) to guide new rater feedback;  
5  
6 this is especially important in a setting where the concept of rater training is novel. In our study,  
7  
8 Zabar's framework was simple and easy to use as evidenced by a decreased level of external  
9  
10 coaching each day.  
11  
12  
13

14  
15 A study strength was the achievement of a level of rater agreement similar to the few  
16  
17 published training course reports for ECEB and HBB. In our participant group, the 'moderate to  
18  
19 good' kappa for the ECEB OSCE was as reported by Kassick and colleagues in Ghana, the only  
20  
21 other ECEB reported study to include in-country evaluators; a regional and national evaluator.<sup>10</sup>  
22  
23 In the HBB OSCE, our findings demonstrated 'fair to moderate' kappa value which was similar to  
24  
25 the 'fair to good' kappa value reported by Reisman and colleagues in Tanzania[15] whose raters  
26  
27 included two external evaluators and one country based evaluator. Comparable studies for kappa  
28  
29 value results for raters scoring the BAB OSCE module are not reported.  
30  
31  
32

33  
34 In training raters, certain challenges were noted. There was unanticipated variance in rater  
35  
36 perceptions of the expected practice standard. Raters were recruited by clinician researchers based  
37  
38 on recollections of which previous participants from recent HBB, ECEB, and BAB trainings had  
39  
40 performed well; no objective strategy had been employed in their selection. This may contribute  
41  
42 to the unanticipated variance in new rater proficiency.  
43  
44

45  
46 Rater trainees were challenged by OSCE items where scores incorporated multi-steps for  
47  
48 their achievement; this was consistent with experiences described by Seto and colleagues who also  
49  
50 identified lower rater agreement for HBB OSCE multi-step items.[16] For example, in our study,  
51  
52 one HBB OSCE 'item' requires the learner to 'prepare the area for delivery'. To achieve a point  
53  
54 and 'pass' this item, the learner must complete all four of: (1) place towels at bedside; (2) place  
55  
56  
57  
58  
59  
60

1  
2  
3 suction at bedside; (3) place a bag and mask at bedside; and (4) place oxytocin at bedside. This  
4  
5 ‘item’ created confusion amongst rater trainees; during mock session review, several participants  
6  
7 had ‘passed’ the mock scenario learner on this item despite not having seen all steps yet having  
8  
9 observed at least one step. To address this gap, we added sub-item tracking boxes; the use of this  
10  
11 strategy warrants further study.  
12  
13

14  
15 Our study was limited by lack of formal training and experience in role-playing by  
16  
17 simulated patients. Our ‘actors’ were not professionally trained (but rather clinicians!) and  
18  
19 scenarios and levels were de novo; ideally, with more resources and time, mock scenarios would  
20  
21 be formally scripted and/or video-captured to optimize standardization. Additionally, time  
22  
23 constraints necessitated working three long days; rater fatigue was likely. This was especially true  
24  
25 for one pregnant rater-trainee who participated for the first two days then arrived with newborn in  
26  
27 hand on Day 3.  
28  
29

## 30 31 **CONCLUSION**

32  
33 Our study shows in rural, Tanzania, training of in-country raters is feasible and effective.  
34  
35 This is the first study of its kind in Africa. We hope our experience encourages program developers  
36  
37 nationally and internationally to scale up in-country rater training. For LMIC simulation-based  
38  
39 training programs to be sustainable, all countries and regions should have their own trained OSCE  
40  
41 raters.  
42  
43

44  
45 Rater training is critical for administering OSCE based learner assessments to maximize  
46  
47 reliability and validity of learner outcomes. Global training programs, including HBB, ECEB and  
48  
49 the BAB need to be confident that OSCE scores truly reflect learner ability, to identify and support  
50  
51 those needing further skill practice. Significant global investments have been made towards  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

maternal newborn health provider training; participants need to leave workshop venues equipped with the skills to save mother and newborn lives.

Confidential: For Review Only

### 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 108024-001** under Mama na MToto programme in rural Tanzania.

Study sponsors had no involvement in study design, collection and analysis of data, interpretation or writing of this manuscript.

### Acknowledgements

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

What is already known:

1. Few HBB, ECEB and BAB study results that report improvements post training in healthcare provider skill report rater agreement
2. There is a gap in our knowledge about the relationship between rater training, rater agreement and participant performance
3. There is a gap in our knowledge the appropriate rater training curriculum for in country raters in LMICs

What this study adds:

1. A conceptual framework for training in country health providers as raters in an LMIC
2. It is possible to achieve moderate rater agreement within country healthcare providers as Raters in an LMIC
3. OSCE checklist multi-step items add complexity and should be adapted to a local context

## 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, 2<sup>nd</sup> edition. 2015. Available from <https://www.aap.org/en-us/advocacy-and-policy/aap-health-initiatives/helping-babies-survive/Pages/Helping-Babies-Breathe.aspx> (Accessed 19 March 2018)
3. Jhpeigo Helping Mothers Survive Training Skills for Health Care Providers, Third Edition: Reference Manual. Editors: Bluestone J, Fowler R, Johnson P, Smith J. Published by Jhpeigo Corporation, USA. 2010. Available: [http://resources.jhpeigo.org/system/files/resources/trainingskills\\_manual\\_0.pdf](http://resources.jhpeigo.org/system/files/resources/trainingskills_manual_0.pdf).
4. Jhpeigo. Helping Mothers Survive Bleeding After Birth Training Package. 2016 Available from <http://reprolineplus.org/resources/HMS>. (Accessed 19 March 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. *International Journal of Gynecology & Obstetrics* 2014;126(3):286-90.
7. Beena D, Kamath-Rayne BD, Thukral A, et al (2018). Helping Babies Breathe, Second Edition: A Model for Strengthening Educational Programs to Increase Global Newborn Survival. *Global Health: Science and Practice*;2018; 6(3): 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;93:287–295.
9. Brucker M. Management of the third stage of labor: an evidence-based approach. *J Midwif Women’s Health*. 2001;46:381-92
10. Kassick M, Chinbuah M, Serpa M, et al. Evaluating a novel neonatal-care assessment tool among trained delivery attendants in a resource-limited setting. *International Journal of Gynecology and Obstetrics* 2018;135(3):285-89.
11. Alwy F, Pembe AB, Hirose A, et al. Effect of the competency based Helping Mothers Survive Bleeding after Birth (HMS BAB) training on maternal morbidity: A cluster randomized trial in 20 districts in Tanzania. *British Medical Journal Global Health* 2019;4(2):e001214.
12. Bishanga DR, Charles J, Tibaijiuka G, et al. Improvement in the active management of the third stage of labor for 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 Practice & Research Clinical Obstetrics and Gynaecology*;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(5):300–08.

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. *International Journal of Gynecology and Obstetrics* 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. *Perspectives on Medical Education* 2015; 4(5):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(9):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. Objective structured clinical examinations. Oxford University Press, Oxford UK. 2013.
20. Van Der Vleuten CPM, Van Luyk SJ, Van Ballegooijen AMJ, et al. Training and experience of examiners. *Med Educ* 1989;23:290-96.
21. Harden, RM. "Revisiting 'Assessment of clinical competence using an objective structured clinical examination (OSCE)'" *Med Educ* 2016;50(4): 376-79.
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(4): 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 Medical Education* 2017;17(71): 1-7.
24. Pugh Vijay John Daniels & Debra. "Twelve tips for developing an OSCE that measures what you want." *Medical Teacher* 2018;40(12):1208-13.
25. Preusche I, Schmidts M, Wagner-Menghin M. 2012. "Twelve tips for designing and implementing a structured rater training in OSCEs." *Medical Teacher* 2012;34(5):368-372.
26. 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." 1-154. Available: <https://dc.sourceafrica.net/documents/118198-Tanzania-Health-Sector-Strategic-Plan-July-2015.html> (Accessed 19 March 2017).
27. Fleiss JL, Cohen J. "The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability." *Educational and Psychological Measurement* 1973;33: 613-619.
28. Zabar S, Krajic Kacher E, Kalet A, et al. 2013. *Objective Structured Clinical Exams- 10 steps to planning and implementing OSCE's and other standardized patient exercises*. Edited by Kachur E, Hanley K. Zabar S. New York: Springer.
29. World Health Organization, OECD, and International Bank for Reconstruction and Development/The World Bank, 2018. 2018. "Delivering quality health services. A global imperative for universal health coverage."
30. 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](https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(18)30386-3/fulltext) (Accessed 6 November 2018).

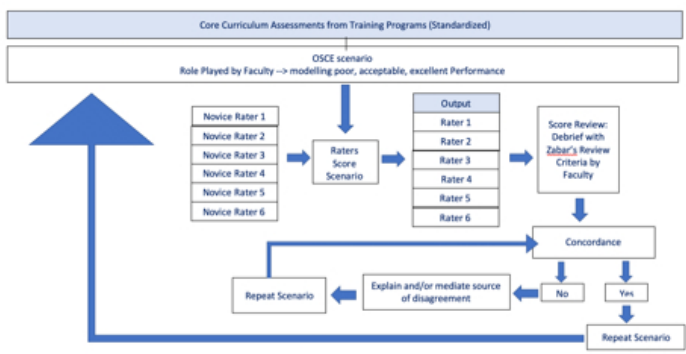
- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60
31. Rowe AK, Labadie G, Jackson D, et al. 2018. "Improving health worker performance an ongoing challenge for meeting the sustainable development goals." *BMJ* 2018 (362): k2813.
  32. Fuller R, Homer M, Pell G, et al. (2017) "Managing extremes of assessor judgment within the OSCE." *Medical Teacher* 2017;39(1):58-66.
  33. Petrusa ER. (2002). 'Clinical Performance Assessments' in Norman G, van der Vleuten C, Newble D (eds) *International Handbook of Reserach in Medical Education* Boston, Kluwer Academic Publishers.673-709.
  34. Reid K, Smallwood D, Collins M, et al. "Taking OSCE examiner training on the road: reaching the masses." *Medical Education Online* 2016;21(1).

Confidential: For Review Only

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

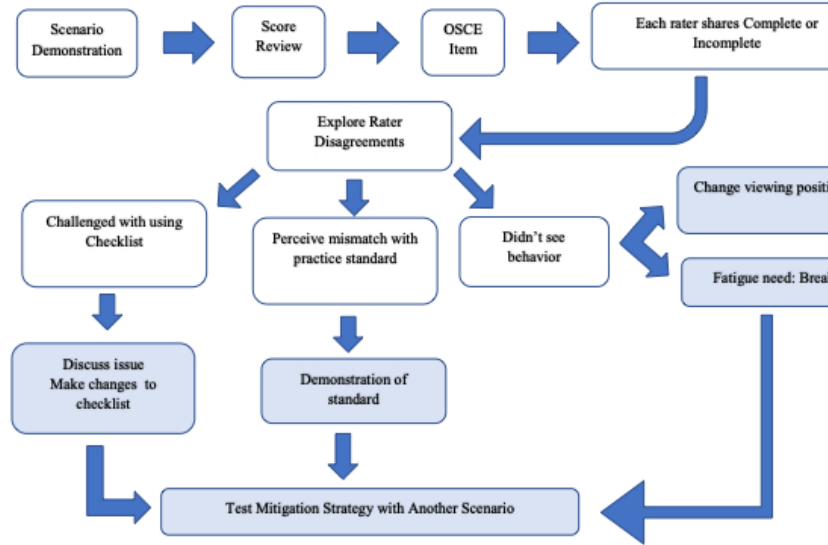


1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



**Figure 1.** This figure provides a visual of the research design we used in this study. All six raters scored each of the 42 role played scenarios depicting poor, average and excellent levels of performance.

215x279mm (72 x 72 DPI)



**Figure 2.** This figure summarizes the conceptual framework we used for score review and feedback. Zabar’s review criteria provides guidance in score review with identification of source of rater agreement and mitigation strategy. This framework reflects the experiential learning cycle starting with experience, an opportunity to reflect (item review), abstract conceptualization (use feedback to rethink experience) and direct experimentation (another opportunity) to apply learning.

215x279mm (72 x 72 DPI)

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

# BMJ Paediatrics Open

## Rater training for standardized assessment of Objective Structured Clinical Exams in rural Tanzania

Journal:	<i>BMJ Paediatrics Open</i>
Manuscript ID	bmjpo-2020-000856.R1
Article Type:	Original research
Date Submitted by the Author:	19-Oct-2020
Complete List of Authors:	Sigalet, Elaine; University of Calgary Cumming School of Medicine, Community Health Sciences Matovelo, Dismas; Catholic University of Health and Allied Sciences Brenner, Jennifer; University of Calgary Cumming School of Medicine, Faculty of Medicine Boniphace, Maendeleo; Catholic University of Health and Allied Sciences Ndaboine, Edgar; Catholic University of Health and Allied Sciences Mwaikasu, Lusako; Catholic University of Health and Allied Sciences Shabani, Girles; Catholic University of Health and Allied Sciences Kabiligi, Julieth; Catholic University of Health and Allied Sciences Mannerfeldt, Jaelene; University of Calgary Cumming School of Medicine, Community Health Sciences Singhal, Nalini; University of Calgary Cumming School of Medicine, Community Health Sciences
Keywords:	Health services research, Resuscitation

SCHOLARONE™  
Manuscripts



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

1 Rater training for standardized assessment of Objective Structured Clinical Exams in rural  
2 Tanzania

3 **Authors:** Elaine Sigalet<sup>1</sup>, Dismas Matovelo<sup>2</sup>, Jennifer L Brenner<sup>1</sup>, Maendeleo Boniphace<sup>2</sup>, Edgar  
4 Ndaboine<sup>2</sup>, Lusako Mwaikasu<sup>2</sup>, Girles Shabani<sup>2</sup>, Julieth Kabiligi<sup>2</sup>, Jaelene Mannerfeldt<sup>1</sup>, Nalini  
5 Singhal<sup>1</sup>

6 **Institution Affiliations:**

7 <sup>1</sup>University of Calgary, Cummings School of Medicine, Alberta  
8 Canada

9 <sup>2</sup>Catholic University of Health & Allied Sciences, Tanzania

10 **Corresponding author:** Elaine Sigalet [E-mail: [elaine.sigalet@gmail.com](mailto:elaine.sigalet@gmail.com)], Department of  
11 Community Health Sciences, University of Calgary Cummings School of Medicine, 3330 Hospital  
12 Drive NW, Calgary, AB, Canada T2N4N1, Tel.+587-4386604

13 **Co-Authors:**

14 **Dismas Matovelo**, Department of Obstetrics and Gynecology, Catholic University of Health &  
15 Allied Sciences, Bugando Medical Center, Mwanza, Tanzania

16 **Jennifer L Brenner**, Department of Pediatrics and Community Health Sciences, Director Global  
17 Maternal Newborn Child Health, University of Calgary Cummings School of Medicine, Calgary,  
18 Alberta Canada

19 **Maendeleo Boniphace**, School of Nursing, Catholic University of Health & Allied Sciences,  
20 Bugando Medical Center, Mwanza, Tanzania

21 **Edgar Ndaboine**, Department of Obstetrics and Gynecology, Catholic University of Health &  
22 Allied Sciences, Bugando Medical Center, Mwanza, Tanzania

23 **Lusako Mwaikasu**, Department of Obstetrics and Gynecology, Catholic University of Health &  
24 Allied Sciences, Bugando Medical Center, Mwanza, Tanzania

25 **Girles Shabani**, Research Coordinator Mama na Mtoto Project, Catholic University of Health &  
26 Allied Sciences, Bugando Medical Center, Mwanza, Tanzania

27 **Julieth Kabiligi**, Department of Pediatrics, Catholic University of Health & Allied Sciences,  
28 Bugando Medical Center, Mwanza, Tanzania

29 **Jaelene Mannerfeldt**, Department of Obstetrics and Gynecology, , University of Calgary  
30 Cummings School of Medicine, Calgary, Alberta Canada

31 **Nalini Singhal**, Department of Neonatology, University of Calgary Cummings School of  
32 Medicine, Calgary, Alberta Canada

33

#### 34 **Contributor Statements:**

35 Elaine Sigalet, Dismas Matovelo, Jennifer L Brenner and Nalini Singhal provided substantial  
36 contributions to the conception and design of the work, drafting and revising the manuscript,  
37 approve the submitted version and agree to be accountable for aspects of the work related to  
38 accuracy or integrity of any part of the work.

39 Girles Shabani contributed substantially to acquisition and analysis of data, revision of manuscript  
40 drafts, approve submitted version and agree to be accountable for all aspects of the work ensuring  
41 questions related to accuracy or integrity are examined and resolved.

42 Maendeleo Boniphace, Edgar Ndaboine, Lusako Mwaikasu, and Julieth Kabiligi contributed  
43 substantially to interpretation of data, revision of manuscript drafts, approve submitted version and  
44 agree to be accountable for all aspects of the work ensuring questions related to accuracy or  
45 integrity are examined and resolved.

1  
2  
3 46 Jaelene Mannerfeldt contributed substantially to conception of work, revision of submitted  
4  
5 47 manuscript, approves submitted manuscript and agrees to be accountable ensuring questions  
6  
7  
8 48 related to accuracy or integrity are examined and resolved.  
9

10 49  
11 50 **Key Words** Simulation Training, Community Child Health, Resuscitation, Mortality, Medical  
12 51 Education  
13  
14 52

15  
16 53 **Word Count: 3352**  
17  
18 54  
19  
20 55  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 **56 Abstract:**

4  
5 **57 OBJECTIVES**

6  
7  
8 58 To describe a simulation-based rater training curriculum for Objective Structured Clinical Exams  
9  
10 59 (OSCEs) for clinician-based training for front line staff caring for mothers and babies in rural  
11  
12 60 Tanzania.

13  
14 **61 BACKGROUND**

15  
16  
17 62 Rater training for OSCE evaluation is widely embraced in high income countries (HIC) but not  
18  
19 63 well described in low and middle-income countries (LMICs). Helping Babies Breathe (HBB),  
20  
21 64 Essential Care for Every Baby (ECEB) and Bleeding after Birth (BAB) are standardized training  
22  
23 65 programs that encourage OSCEs evaluations. Studies examining the reliability of assessments are  
24  
25 66 rare.

26  
27 **67 METHODS**

28  
29  
30 68 Training of raters occurred over three days. Raters scored selected OSCEs role played using  
31  
32 69 standardized learners and low fidelity mannikins, assigning proficiency levels *a priori*.  
33  
34 70 Researchers used Zabbar's criteria to critique rater agreement and mitigate measurement error  
35  
36 71 during score review. Descriptive statistics, Fleiss' kappa and field notes were used to describe  
37  
38 72 results.

39  
40 **73 RESULTS**

41  
42  
43 74 Six healthcare providers scored 42 training scenarios. There was moderate rater agreement across  
44  
45 75 all OSCEs ( $\kappa=0.508$ ). Kappa values increased with HBB ( $\kappa=0.28$  to  $0.48$ ), and ECEB ( $\kappa=0.42$  to  
46  
47 76  $0.77$ ) by Day 3 of training but not with BAB ( $\kappa=0.58$  to  $0.33$ ). Raters identified average  
48  
49 77 proficiency 50% of the time.

50  
51 **78 CONCLUSION**



1  
2  
3 79 Our study shows training of in-country raters resulted in the discernment of acceptable proficiency  
4  
5 80 50% of the time, despite moderate rater agreement. Rater training is critical to ensure that the  
6  
7  
8 81 potential of training programs translates to improved outcomes for mothers and babies; more  
9  
10 82 research into the concepts and training for discernment of competence in this setting is necessary.  
11  
12  
13 83

Confidential: For Review Only

## 84 BACKGROUND

85 Helping Babies Breathe (HBB) and Essential Care for Every Baby (ECEB), from the Helping  
86 Babies Survive Program[1,2] and the Bleeding after Birth (BAB) from the Helping Mothers  
87 Survive (HMS) program[3,4] are examples of standardized health provider training programs  
88 designed by expert clinicians and educators from high income countries (HIC) with input from  
89 low and middle income countries (LMICs) for use in LMICs. The HBB training course reviews  
90 skills related to newborn resuscitation; ECEB focuses on newborn routine care and danger sign  
91 identification; BAB reviews management of maternal hemorrhage. All three courses and others  
92 in the HMS, HBS series, use low-fidelity mannequins, hands-on simulation practice of common  
93 case scenarios and emphasize compliance with algorithm-based ‘Action Plans’. Course content  
94 addresses common gaps that lead to some of the highest sources of global maternal[5,6] and  
95 newborn mortality.[1,2]

96 The competence of participants in these courses are frequently assessed using Objective  
97 Structured Clinical Exams (OSCEs). A number of studies in a variety of LMIC settings have  
98 demonstrated improvements in provider competency managing relevant obstetric and neonatal  
99 cases post training.[6-16] However, few of these studies provide details of assessor training, or  
100 the reliability of the OSCE assessments.[10,15,16] Furthermore, only one study used in-country  
101 OSCE raters;[15] others have relied on external (from outside the country of study) development  
102 and academic partners serving in rater roles.[9,16] Training of raters to serve as OSCEs assessors  
103 is widely embraced in HIC,[17-25] but rater training has not been well described in LMICs.  
104 Reisman and colleagues refer to standardized OSCE training but do not report details.[15] Formal  
105 pre-OSCE training for assessors aims to minimise sources of measurement error,[17-25]  
106 increasing confidence that a participant’s OSCE score truly reflects their competence. With OSCE

1  
2  
3 107 administration, sources of error can arise from the OSCE structure and/or rater  
4  
5 108 objectivity.[17,19,22,25] Facilitator materials for HBB, ECEB and BAB courses provide clear  
6  
7  
8 109 guidelines to minimise measurement error with the OSCE administration. For example, Jhpeigo  
9  
10 110 provides information on quality assessment[3] for their HMS training series, but there are no  
11  
12 111 guidelines for training OSCE raters or evaluating rater agreement. The purpose of our study was  
13  
14 112 to describe a simulation based OSCE rater training curriculum and assessment of subsequent levels  
15  
16 113 of rater agreement with administration of OSCEs in rural Tanzania using locally trained healthcare  
17  
18 114 providers as raters.

## 115 **METHOD**

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

### 121 **Patient and Public Involvement**

122 Patients were not involved in this study.

### 123 **Setting**

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

### 126 **Participants**

127 Raters were recruited from clinical staff practising in the rural health facilities in the district where  
128 training was to occur. Selection was based on their demonstrated proficiency in previous Newborn  
129 Maternal training workshops conducted in the previous year. All trainees were clinically active in

1  
2  
3 130 their health facility settings. All selected participants provided informed consent to be involved in  
4  
5 131 the study. Rater characteristics and Rater OSCE scores for each OSCE scenario were collated  
6  
7  
8 132 under a master tracking number to ensure rater anonymity. Following three days of rater training,  
9  
10 133 participants were involved as raters for OSCE evaluations to assess workshop learners pre and post  
11  
12 134 training, at 6 and at 12 months. The rater training curriculum was led by a team comprised of  
13  
14 135 clinician researchers from Catholic University of Health and Allied Sciences (CUHAS) and  
15  
16 136 University of Calgary.

### 19 137 **Design**

20  
21 138 This study used a descriptive study design (Figure 1). Raters attended rater training prior to any  
22  
23 139 formal scoring of workshop participants. Categorical levels of proficiency (poor, acceptable and  
24  
25 140 excellent) (decided a priori) were role modelled by clinician research team members for each  
26  
27 141 OSCE each day to create a mock scoring context. All six raters observed and scored the exact  
28  
29 142 same scenario at the same time, making judgements about observed behaviors independent of  
30  
31 143 discussion with each other. Scores were collected and then reviewed with the raters; areas of  
32  
33 144 disagreement were explored, using an inquiry approach and direct feedback in debriefing.  
34  
35 145 Zabar's review criteria and mitigation strategies was used as the framework for both the reviews  
36  
37 146 and refining methodology. Categorical levels of proficiency that challenged rater agreement  
38  
39 147 were repeated. Checklists were collected and collated on an MS Excel spreadsheet on a research  
40  
41 148 dedicated computer. Field notes were used to track challenges. SPSS version 26 was used to  
42  
43 149 analyse rater data. Descriptive statistics were used to provide information about mock scoring  
44  
45 150 and rater's abilities to identify the three categorical levels of proficiency. All raw scores  
46  
47 151 indicating excellent levels of proficiency (Table 2) were also analyzed as acceptable (Table 3) to  
48  
49 152 align with training program guidelines; two categories of proficiency. Fleiss' Kappa with  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 153 standard error was calculated to provide information about the level of rater agreement.[26]

4  
5 154 Kappa values of <0.20, 0.21-0.40, 0.41-0.60 and 0.61-0.80 and 0.81 to 1.00 are considered poor,  
6  
7 155 fair, moderate, good, and very good respectively.[26]

### 10 156 **Evaluation Tools**

11  
12 157 The OSCEs used were drawn from training program materials.[1-4] There were 24 pass/fail items  
13  
14 158 on the HBB OSCE, 15 items on the ECEB OSCE and 14 items on the BAB OSCE. All raters were  
15  
16 159 familiar with the OSCE checklists and relevant training course content as they had recently  
17  
18 160 participated in the same courses themselves as learners. Poor proficiency, often referred to as ‘red’  
19  
20 161 in reported studies was identified by a score of <71%; 0-17, 0-10, and 0-9 on the HBB, ECEB and  
21  
22 162 BAB OSCE, respectively. Learner scores >70% identified an ‘acceptable’ level of proficiency or  
23  
24 163 ‘green’ in reported studies; >17, >10, & >9 on HBB, ECEB and BAB respectively.[1-4] The  
25  
26 164 Research team added a third category, a candidate’s score of >22, >13 and >12 identified excellent  
27  
28 165 proficiency for HBB, ECEB and BAB OSCEs, respectively. To standardize the proficiency level in  
29  
30 166 a scenario, a priori the researchers used the clinical consequences of an action to inform the  
31  
32 167 scoring, which was then used to plan the actions role played in the scenario.

### 37 168 **The Rater Curriculum**

38  
39 169 The conceptual framework (Figure 2) and Zabar’s review criteria[27] provides details about  
40  
41 170 elements of the curriculum and the iterative nature of the training process. Three physical OSCE  
42  
43 171 stations were set up to facilitate learner transition between each testing station. Checklists were  
44  
45 172 reviewed prior to scoring practice Day 1 of training to ensure raters were familiar with OSCE  
46  
47 173 items and how to use the checklist in scoring. Raters observed a scenario, with a predetermined  
48  
49 174 level of proficiency. Training of raters occurred in the score review, with faculty leading  
50  
51 175 discussions to discern the underlying ideas or concepts which may have led to the disagreement.  
52  
53  
54  
55  
56  
57  
58  
59  
60

176 Raters learned about potential sources of error in the discussion of rater disagreements in score  
 177 review. Faculty discussed the importance of mitigating these sources of error to improve score  
 178 reliability. Scenarios with disagreement on two or more items were repeated.

179

## 180 RESULTS

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

185 **Table 1.** Kappa values

186

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

187

188 The time needed for each OSCE station with score review was longer for average proficiency  
 189 levels (30-40 minutes) when compared to 'excellent' and 'poor' proficiency levels (15-20  
 190 minutes). Fleiss' Kappa values (Table 1) showed that there was a moderate level of rater  
 191 agreement in identifying 'poor' and 'acceptable' proficiency across all OSCEs ( $\kappa=0.51$ ). Kappa  
 192 values improved over the three days moving from 'fair' to 'moderate' for the HBB OSCE and

193 'moderate' to 'good' for the ECEB OSCE. The kappa value for BAB was 'moderate' Day 1 but  
 194 decreased to 'fair' Day 2 and Day 3. Information about rater abilities to correctly identify  
 195 proficiency levels is described in Table 2 and 3.

196

**Table 2.** 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	Rater5	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%)

197

198

199

**Table 3.** Proficiency level identification: (average and excellent categories combined) for training program 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%)

200

201 Raters were more accurate in identifying 'poor' and 'excellent' compared to average. Raters  
 202 identified average proficiency approximately 50% of the time (Table 2 and 3). Information  
 203 detailing challenges from field notes are presented in Table 4.

204



205

**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 multi-step OSCE items	Item 1. Prepares area for delivery  Item 2. Equipment preparation Item 3. Hand washing Item 5. Removes wet clothes Item 24. Communication and teaching	Item 7: Improves thermal care  Item 8: Identifying danger signs Advanced care classification Item 10. Medication calculation and administration	Item 7. Controlled cord traction counter pressure Item 12. Determining Postpartum hemorrhage
OSCE English Words		Hypothermia	Hypertension
Actions without verbalizing		Warming baby	

206

**DISCUSSION**

208 This study describes an OSCE rater training curriculum and presents evaluation of the  
 209 curriculum showing levels of rater agreement for HBB, ECEB and BAB training courses in an  
 210 LMIC. Quality rater training and subsequent reliability analysis is especially important in LMIC  
 211 context because of the limited quality assurance monitoring patient safety in the system and  
 212 resources.[28-31] Our results suggest that the moderate levels of rater agreement, coupled by  
 213 notable challenges in discriminating ‘acceptable’ performance, exposes a potential for either

1  
2  
3 214 overestimating or underestimating competence. This has consequences for the individual, the  
4  
5 215 training program, and the system. The challenge incurred in discriminating between borderline  
6  
7 216 performance is not isolated to an LMIC context but reported universally.[32-34] With  
8  
9  
10 217 overestimation of competence, training programs may have passed clinicians who may need  
11  
12 218 more training to provide safe care on the frontline. The problem of accurate discrimination of  
13  
14 219 competency affects resource utilization: with underestimation of competence, training programs  
15  
16 220 may be directing the limited resources to clinicians who do not need extra training. Furthermore,  
17  
18 221 frontline staff frequently work short staffed when someone is away at training, so unnecessary  
19  
20 222 remediation training may exacerbate staff overload.[28-31]  
21  
22  
23

24 223 In the majority of HBB, ECEB and BAB training program reports, validation of  
25  
26 224 improved care-giver competency is determined by comparing pre and post training OSCE scores.  
27  
28 225 Our results suggest that the existing reports describing a moderate IRR may be misleading  
29  
30 226 without further validation of the accuracy of rater discernment of acceptable proficiency.  
31  
32  
33 227 [10,15,16] Our raters achieved moderate rater agreement yet discernment of acceptable  
34  
35 228 proficiency, which is the pass criterion in these training programs, was approximately 50%.  
36  
37 229 Based on our findings we would suggest including both measures of validation) Considering  
38  
39 230 contexts with limited resources, it may be helpful to implement a further strategy such a global  
40  
41 231 rating scale, which is common practice in the developed world[17-19,22-25] to provide another  
42  
43 232 method of validation of participant competence.[35] A global rating scale allows the rater to  
44  
45 233 evaluate how well a learner performs on a scale of 1 to 5, with 5 reflecting the highest level of  
46  
47 234 competence.[35] More than one method of validation creates more certainty that results are an  
48  
49 235 accurate reflection of participant competence and/or training program efficacy.[35] With the  
50  
51 236 continued high reports of maternal and neonatal mortality, it is important to be confident that  
52  
53  
54  
55  
56  
57  
58  
59  
60

237 these training programs are accurate in identifying and supporting clinicians who may not be  
238 providing safe care on the frontline. Based on our findings we would suggest including both  
239 measures of validation.

240  
241 The guidelines for OSCE rater training used in this study were based on  
242 recommendations from HIC rater training experiences; these are challenging to implement in an  
243 LMIC context. Globally, good practice is for OSCE raters to have relevant content expertise, be  
244 well orientated to the OSCE checklist and use a validated rating scale.[22-26] Although we  
245 strived for this, we had a limited pool of potential raters; this may have affected the challenges  
246 we noted in rater perceptions of the expected practice standard. Raters were recruited by  
247 clinician researchers based on recollections of which previous participants from recent HBB,  
248 ECEB, and BAB trainings had performed well; no objective strategy was employed in their  
249 selection. This was the reason in country faculty inserted a third categorical level of proficiency;  
250 excellent. They wanted an objective strategy to identify content experts as the future raters for  
251 such training programs. A quality rater training curriculum includes standardized mock  
252 scenarios where raters practise and score a variety of expected learner proficiency levels. In our  
253 study, this was one of the greatest challenges. Research clinicians role playing scenarios Day 1  
254 were challenged in demonstrating poor proficiency. In discussion, they shared they didn't want  
255 participants to think they were not experts in the field. The inclusion of scripted and video  
256 capture of proficiency levels may lessen this tension and inconsistency in role play. In a limited  
257 resource setting this is challenging to develop and implement. Despite this, the level of rater  
258 agreement improved over the three training days for both HBB and ECEB. The fall-off in rater  
259 agreement for BAB Day 2 and 3 was unexpected but may be in part related to the timing of these  
260 scenarios; they were the last role plays of the day and rater fatigue may have played a role.

1  
2  
3 261 A solid rater curriculum incorporates a framework such as Zabar's (Figure 2) to guide rater  
4  
5 262 feedback; this is especially important in a setting where the concept of rater training is novel.[27]  
6  
7  
8 263 In our study, Zabar's framework was simple and easy to use as evidenced by a decreased level of  
9  
10 264 external coaching each day.

11  
12 265 A study strength was the achievement of a level of rater agreement similar to the few  
13  
14 266 published training course reports for ECEB and HBB. In our participant group, the 'moderate to  
15  
16 267 good' kappa for the ECEB OSCE was as reported by Kassick and colleagues in Ghana, the only  
17  
18 268 other ECEB reported study to include in-country evaluators; a regional and national  
19  
20 269 evaluator.[10] In the HBB OSCE, our findings demonstrated 'fair to moderate' kappa value  
21  
22 270 which was similar to the 'fair to good' kappa value reported by Reisman and colleagues in  
23  
24 271 Tanzania[15] whose raters included two external evaluators and one country based evaluator.  
25  
26 272 Comparable studies for kappa value results for raters scoring the BAB OSCE module are not  
27  
28 273 reported. The achievement of comparable IRR to the studies using in country and external  
29  
30 274 partners provides support for the rater training curriculum, yet the inability to accurately discern  
31  
32 275 acceptable proficiency (pass criteria) is concerning. To gain further insight into the relationship  
33  
34 276 between faculty role play and the inability to discern acceptable proficiency, we plan to script the  
35  
36 277 acceptable proficiency level for each OSCE, coach faculty in the role play, and repeat the  
37  
38 278 curriculum and analysis.

39  
40 279 Rater trainees were challenged by OSCE items where scores incorporated multi-steps for  
41  
42 280 their achievement; this was consistent with experiences described by Seto and colleagues who also  
43  
44 281 identified lower rater agreement for HBB OSCE multi-step items.[16] For example, in our study,  
45  
46 282 one HBB OSCE 'item' requires the learner to 'prepare the area for delivery'. To achieve a point  
47  
48 283 and 'pass' this item, the learner must complete all four of: (1) place towels at bedside; (2) place  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 284 suction at bedside; (3) place a bag and mask at bedside; and (4) place oxytocin at bedside. This  
4  
5 285 'item' created confusion amongst rater trainees; during mock session review, several participants  
6  
7  
8 286 had 'passed' the mock scenario learner on this item despite not having seen all steps yet having  
9  
10 287 observed at least one step. To address this gap, we added sub-item tracking boxes when this  
11  
12 288 challenge was identified Day 1; the use of this strategy warrants further study.

13  
14  
15 289 Our study was limited by lack of formal training and experience in role-playing by  
16  
17 290 simulated learners. Our 'actors' were not professionally trained (but rather research clinicians!)  
18  
19 291 and scenarios and levels were de novo; ideally, with more resources and time, mock scenarios  
20  
21 292 would be formally scripted and/or video-captured to optimize standardization. Additionally, time  
22  
23 293 constraints necessitated working three long days; rater fatigue was likely. This was especially true  
24  
25 294 for one pregnant rater-trainee who participated for the first two days then arrived with newborn in  
26  
27 295 hand on Day 3. Our results may have limitations in generalisability but do provide context and  
28  
29 296 learning for others interested in developing a rater training curriculum in a low resource setting.

## 30 31 32 33 297 34 35 298 **CONCLUSION**

36  
37 299 Our results show that rater training in an LMIC setting is critical for administering OSCE based  
38  
39 300 learner assessments. Clinician everywhere need ongoing training, but to optimize learning and  
40  
41 301 then translate this to improved outcomes for mothers and babies, this training must be informed  
42  
43 302 by truly objective evaluations. Our study shows in rural, Tanzania, training of in-country raters is  
44  
45 303 possible and can lead to an IRR which is similar to previous studies. Improved standardization  
46  
47 304 and attention to the relationships between IRR and the accurate discernment of participant  
48  
49 305 performance would provide insight into needed modifications, which in turn may lead to greater  
50  
51 306 accuracy in rating competence. More research is warranted. Global training programs, including  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 307 HBB, ECEB and the BAB need to be confident that OSCE scores truly reflect learner ability, to  
4  
5 308 identify and support those needing further skill practice. Significant global investments have been  
6  
7 309 made towards maternal newborn health provider training; participants need to leave workshop  
8  
9 310 venues equipped with the skills to save mother and newborn lives. We hope this experience  
10  
11 311 encourages program developers nationally and internationally to scale up in-country rater training.  
12  
13 312 For LMIC simulation-based training programs to be sustainable, all countries and regions should  
14  
15 313 have their own trained OSCE raters.  
16  
17  
18  
19 314  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 **316 Funding**  
4

5 317 “This work was supported by a grant from the Innovating for Maternal and Child Health in Africa  
6 318 (IMCHA) initiative- a partnership of Global Affairs Canada (GAC), the Canadian Institutes of Health  
7 319 Research (CIHR) and Canada’s International Development Research Centre (IDRC), **Grant number**  
8 320 **108024-001** under Mama na MToto programme in rural Tanzania.  
9 321

10 322 Study sponsors had no involvement in study design, collection and analysis of data, interpretation or  
11 323 writing of this manuscript.  
12 324

13 325 **Acknowledgements**  
14  
15

16 326 Healthcare workers from Misungwi District who served as raters for this training program.  
17 327  
18 328  
19 329  
20 330

21 331 What is already known:

- 22 332 1. Few HBB, ECEB and BAB study results that report improvements post training in  
23 333 healthcare provider skill report rater agreement  
24 334 2. There is a gap in our knowledge about the relationship between rater training, rater  
25 335 agreement and participant performance  
26 336 3. There is a gap in our knowledge the appropriate rater training curriculum for in country  
27 337 raters in LMICs

28 338 What this study adds:

- 29 339 1. A conceptual framework for training in country health providers as raters in an LMIC  
30 340 2. It is possible to achieve moderate rater agreement within country healthcare providers as  
31 341 Raters in an LMIC  
32 342 3. OSCE checklist multi-step items add complexity and should be adapted to a local context  
33 343  
34 344  
35 345  
36 346  
37 347  
38 348  
39 349  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## 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, 2<sup>nd</sup> edition. 2015. Available from <https://www.aap.org/en-us/advocacy-and-policy/aap-health-initiatives/helping-babies-survive/Pages/Helping-Babies-Breathe.aspx> (Accessed 19 March 2018)
3. Jhpeigo Helping Mothers Survive Training Skills for Health Care Providers, Third Edition: Reference Manual. Editors: Bluestone J, Fowler R, Johnson P, Smith J. Published by Jhpeigo Corporation, USA. 2010. Available: [http://resources.jhpiego.org/system/files/resources/trainingskills\\_manual\\_0.pdf](http://resources.jhpiego.org/system/files/resources/trainingskills_manual_0.pdf).
4. Jhpeigo. Helping Mothers Survive Bleeding After Birth Training Package. 2016 Available from <http://reprolineplus.org/resources/HMS>. (Accessed 19 March 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. *International Journal of Gynecology & Obstetrics* 2014;126(3):286-90.
7. Beena D, Kamath-Rayne BD, Thukral A, et al (2018). Helping Babies Breathe, Second Edition: A Model for Strengthening Educational Programs to Increase Global Newborn Survival. *Global Health: Science and Practice*;2018; 6(3): 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;93:287–295.
9. Brucker M. Management of the third stage of labor: an evidence-based approach. *J Midwif Women’s Health*. 2001;46:381-92
10. Kassick M, Chinbuah M, Serpa M, et al. Evaluating a novel neonatal-care assessment tool among trained delivery attendants in a resource-limited setting. *International Journal of Gynecology and Obstetrics* 2018;135(3):285-89.
11. Alwy F, Pembe AB, Hirose A, et al. Effect of the competency based Helping Mothers Survive Bleeding after Birth (HMS BAB) training on maternal morbidity: A cluster randomized trial in 20 districts in Tanzania. *British Medical Journal Global Health* 2019;4(2):e001214.
12. Bishanga DR, Charles J, Tibaijiuka G, et al. Improvement in the active management of the third stage of labor for 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 Practice & Research Clinical Obstetrics and Gynaecology*;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(5):300–08.



- 1  
2  
3 396  
4 397  
5 398  
6  
7 399  
8 400  
9 401  
10 402  
11 403  
12 404  
13  
14 405  
15 406  
16 407  
17 408  
18 409  
19 410  
20 411  
21  
22 412  
23 413  
24 414  
25 415  
26 416  
27 417  
28 418  
29  
30 419  
31 420  
32 421  
33 422  
34 423  
35 424  
36 425  
37  
38 426  
39 427  
40 428  
41 429  
42 430  
43 431  
44 432  
45 433  
46 434  
47 435  
48 436  
49 437  
50 438  
51 439  
52 440  
53 441  
54  
55  
56  
57  
58  
59  
60
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. *International Journal of Gynecology and Obstetrics* 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. *Perspectives on Medical Education* 2015; 4(5):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(9):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. Objective structured clinical examinations. Oxford University Press, Oxford UK. 2013.
  20. Van Der Vleuten CPM, Van Luyk SJ, Van Ballegooijen AMJ, et al. Training and experience of examiners. *Med Educ* 1989;23:290–96.
  21. Harden, RM. "Revisiting 'Assessment of clinical competence using an objective structured clinical examination (OSCE)'" *Med Educ* 2016;50(4): 376–79.
  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(4): 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 Medical Education* 2017;17(71): 1-7.
  24. Pugh Vijay John Daniels & Debra. "Twelve tips for developing an OSCE that measures what you want." *Medical Teacher* 2018;40(12):1208-13.
  25. Preusche I, Schmidts M, Wagner-Menghin M. 2012. "Twelve tips for designing and implementing a structured rater training in OSCEs." *Medical Teacher* 2012;34(5):368-372.
  26. Fleiss JL, Cohen J. "The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability." *Educational and Psychological Measurement* 1973;33: 613-619.
  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*. 2013; Edited by Kachur E, Hanley K. Zabar S. New York: Springer.
  28. 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." 1-154. Available: <https://dc.sourceafrica.net/documents/118198-Tanzania-Health-Sector-Strategic-Plan-July-2015.html> (Accessed 19 March 2017).
  29. World Health Organization, OECD, and International Bank for Reconstruction and Development/The World Bank, 2018. "Delivering quality health services. A global imperative for universal health coverage."
  30. 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](https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(18)30386-3/fulltext) (Accessed 6 November 2018).
  31. 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.

- 1  
2  
3 442 32. Fuller R, Homer M, Pell G, et al. "Managing extremes of assessor judgment within the  
4 443 OSCE." *Medical Teacher* 2017;39(1):58-66.  
5 444 33. Petrusa ER. 'Clinical Performance Assessments' in Norman G, van der Vleuten C, Newble  
6 445 D (eds) *International Handbook of Reserach in Medical Education,2002*; Boston, Kluwer  
7 446 Academic Publishers.673-709.  
8 447 34. Reid K, Smallwood D, Collins M, et al. "Taking OSCE examiner training on the road:  
9 448 reaching the masses." *Medical Education Online* 2016;21(1).  
10 449 35. Ilgen, Jonathan S, et al. "A Systematic Review of Validity Evidence for Checklists versus  
11 450 Global Rating Scales in Simulation-Based Assessment." *Medical Education*, 2015,  
12 451 49(2):161–173.  
13 452  
14 453  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Confidential: For Review Only

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

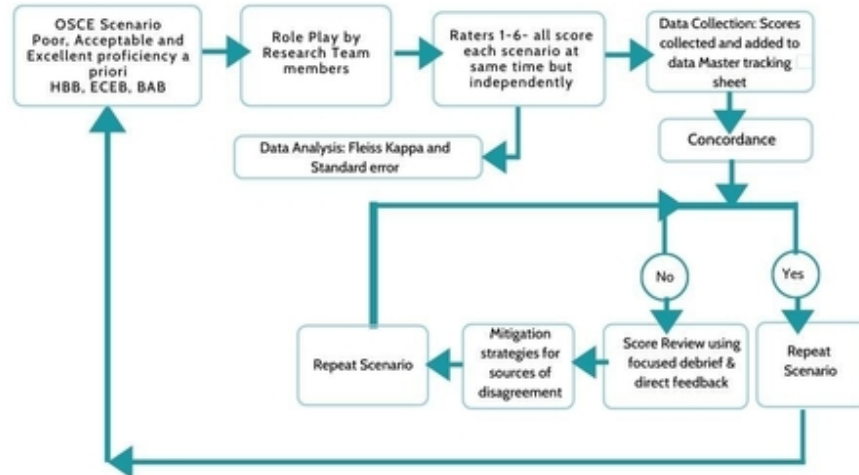


Figure 1. This figure provides a visual of the research design we used in the study each day. All six raters scored all 42 of the role played scenarios with proficiency determined a priori. Raters participated in 42 debrief sessions over the three days.

43x32mm (300 x 300 DPI)

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

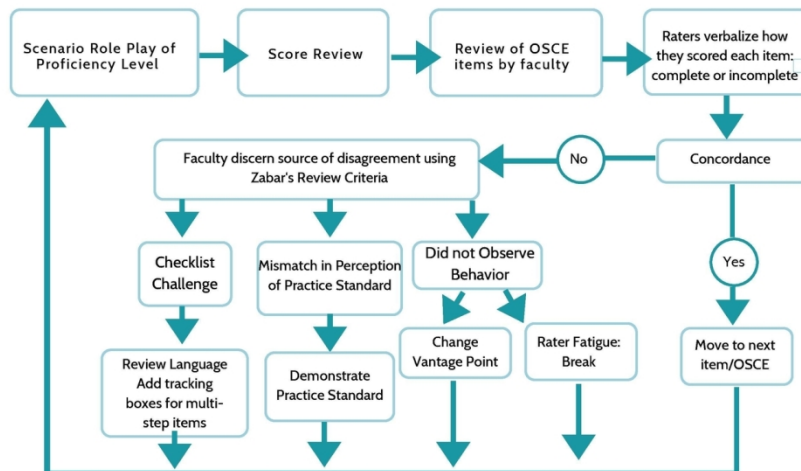


Figure 2. This figure provides a visual of the Conceptual framework used to improve the level of rater agreement.

135x101mm (300 x 300 DPI)

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

# BMJ Paediatrics Open

## Rater training for standardized assessment of Objective Structured Clinical Exams in rural Tanzania

Journal:	<i>BMJ Paediatrics Open</i>
Manuscript ID	bmjpo-2020-000856.R2
Article Type:	Original research
Date Submitted by the Author:	16-Nov-2020
Complete List of Authors:	Sigalet, Elaine; University of Calgary Cumming School of Medicine, Community Health Sciences Matovelo, Dismas; Catholic University of Health and Allied Sciences Brenner, Jennifer; University of Calgary Cumming School of Medicine, Faculty of Medicine Boniphace, Maendeleo; Catholic University of Health and Allied Sciences Ndaboine, Edgar; Catholic University of Health and Allied Sciences Mwaikasu, Lusako; Catholic University of Health and Allied Sciences Shabani, Girles; Catholic University of Health and Allied Sciences Kabiligi, Julieth; Catholic University of Health and Allied Sciences Mannerfeldt, Jaelene; University of Calgary Cumming School of Medicine, Community Health Sciences Singhal, Nalini; University of Calgary Cumming School of Medicine, Community Health Sciences
Keywords:	Health services research, Resuscitation

SCHOLARONE™  
Manuscripts



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

1 Rater training for standardized assessment of Objective Structured Clinical Exams in rural  
2 Tanzania

3 **Authors:** Elaine Sigalet<sup>1</sup>, Dismas Matovelo<sup>2</sup>, Jennifer L Brenner<sup>1</sup>, Maendeleo Boniphace<sup>2</sup>, Edgar  
4 Ndaboine<sup>2</sup>, Lusako Mwaikasu<sup>2</sup>, Girles Shabani<sup>2</sup>, Julieth Kabiligi<sup>2</sup>, Jaelene Mannerfeldt<sup>1</sup>, Nalini  
5 Singhal<sup>1</sup>

6 **Institution Affiliations:**

7 <sup>1</sup>University of Calgary, Cummings School of Medicine, Alberta  
8 Canada

9 <sup>2</sup>Catholic University of Health & Allied Sciences, Tanzania

10 **Corresponding author:** Elaine Sigalet [E-mail: [elaine.sigalet@gmail.com](mailto:elaine.sigalet@gmail.com)], Department of  
11 Community Health Sciences, University of Calgary Cummings School of Medicine, 3330 Hospital  
12 Drive NW, Calgary, AB, Canada T2N4N1, Tel.+587-4386604

13 **Co-Authors:**

14 **Dismas Matovelo**, Department of Obstetrics and Gynecology, Catholic University of Health &  
15 Allied Sciences, Bugando Medical Center, Mwanza, Tanzania

16 **Jennifer L Brenner**, Department of Pediatrics and Community Health Sciences, Director Global  
17 Maternal Newborn Child Health, University of Calgary Cummings School of Medicine, Calgary,  
18 Alberta Canada

19 **Maendeleo Boniphace**, School of Nursing, Catholic University of Health & Allied Sciences,  
20 Bugando Medical Center, Mwanza, Tanzania

21 **Edgar Ndaboine**, Department of Obstetrics and Gynecology, Catholic University of Health &  
22 Allied Sciences, Bugando Medical Center, Mwanza, Tanzania



23 **Lusako Mwaikasu**, Department of Obstetrics and Gynecology, Catholic University of Health &  
24 Allied Sciences, Bugando Medical Center, Mwanza, Tanzania

25 **Girles Shabani**, Research Coordinator Mama na Mtoto Project, Catholic University of Health &  
26 Allied Sciences, Bugando Medical Center, Mwanza, Tanzania

27 **Julieth Kabiligi**, Department of Pediatrics, Catholic University of Health & Allied Sciences,  
28 Bugando Medical Center, Mwanza, Tanzania

29 **Jaelene Mannerfeldt**, Department of Obstetrics and Gynecology, , University of Calgary  
30 Cummings School of Medicine, Calgary, Alberta Canada

31 **Nalini Singhal**, Department of Neonatology, University of Calgary Cummings School of  
32 Medicine, Calgary, Alberta Canada

33

#### 34 **Contributor Statements:**

35 Elaine Sigalet, Dismas Matovelo, Jennifer L Brenner and Nalini Singhal provided substantial  
36 contributions to the conception and design of the work, drafting and revising the manuscript,  
37 approve the submitted version and agree to be accountable for aspects of the work related to  
38 accuracy or integrity of any part of the work.

39 Maendeleo Boniphace, Edgar Ndaboine, Lusako Mwaikasu, and Julieth Kabiligi contributed  
40 substantially to interpretation of data, revision of manuscript drafts, approve submitted version and  
41 agree to be accountable for all aspects of the work ensuring questions related to accuracy or  
42 integrity are examined and resolved.

43 Girles Shabani contributed substantially to acquisition and analysis of data, revision of manuscript  
44 drafts, approve submitted version and agree to be accountable for all aspects of the work ensuring  
45 ensuring questions related to accuracy or integrity are examined and resolved.

1  
2  
3 46 Jaelene Mannerfeldt contributed substantially to conception of work, revision of submitted  
4  
5 47 manuscript, approves submitted manuscript and agrees to be accountable ensuring questions  
6  
7  
8 48 related to accuracy or integrity are examined and resolved.  
9

10 49 **Key Words** Simulation Training, Community Child Health, Resuscitation, Mortality, Medical  
11 50 Education  
12  
13 51

14  
15 52 **Word Count: 2928**  
16  
17 53  
18  
19 54  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 **55 Abstract:**

4  
5 **56 OBJECTIVES**

6  
7  
8 57 To describe a simulation based rater training curriculum for Objective Structured Clinical Exams  
9  
10 58 (OSCEs) for clinician based training for front line staff caring for mothers and babies in rural  
11  
12 59 Tanzania.

13  
14 **60 BACKGROUND**

15  
16  
17 61 Rater training for OSCE evaluation is widely embraced in high income countries but not well  
18  
19 62 described in low and middle-income countries. Helping Babies Breathe, Essential Care for Every  
20  
21 63 Baby and Bleeding after Birth are standardized training programs that encourage OSCEs  
22  
23 64 evaluations. Studies examining the reliability of assessments are rare.

24  
25 **65 METHODS**

26  
27  
28 66 Training of raters occurred over three days. Raters scored selected OSCEs role played using  
29  
30 67 standardized learners and low fidelity mannikins, assigning proficiency levels *a priori*.  
31  
32 68 Researchers used Zabar's criteria to critique rater agreement and mitigate measurement error  
33  
34 69 during score review. Descriptive statistics, Fleiss' kappa and field notes were used to describe  
35  
36 70 results.

37  
38 **71 RESULTS**

39  
40  
41 72 Six healthcare providers scored 42 training scenarios. There was moderate rater agreement across  
42  
43 73 all OSCEs ( $\kappa=0.508$ ). Kappa values increased with Helping Babies Breathe ( $\kappa=0.28$  to 0.48), and  
44  
45 74 Essential Care for Every Baby ( $\kappa=0.42$  to 0.77) by Day 3 of training but not with Bleeding after  
46  
47 75 Birth ( $\kappa=0.58$  to 0.33). Raters identified average proficiency 50% of the time.

48  
49 **76 CONCLUSION**

1  
2  
3 77 Our study shows that the in-country raters in this study had a hard time identifying average  
4  
5 78 performance despite moderate rater agreement. Rater training is critical to ensure that the  
6  
7 79 potential of training programs translates to improved outcomes for mothers and babies; more  
8  
9  
10 80 research into the concepts and training for discernment of competence in this setting is  
11  
12 81 necessary.  
13  
14  
15 82

Confidential: For Review Only

## 83 BACKGROUND

84 Helping Babies Breathe (HBB) and Essential Care for Every Baby (ECEB), from the Helping  
85 Babies Survive Program[1,2] and the Bleeding after Birth (BAB) from the Helping Mothers  
86 Survive (HMS) program[3,4] are examples of standardized health provider training programs  
87 designed by expert clinicians and educators from high income countries (HIC) with input from  
88 low and middle income countries (LMICs) for use in LMICs. The HBB training course reviews  
89 skills related to newborn resuscitation; ECEB focuses on newborn routine care and danger sign  
90 identification; BAB reviews management of maternal hemorrhage. All three courses and others  
91 in the HMS, HBS series, use low-fidelity mannequins, hands-on simulation practice of common  
92 case scenarios and emphasize compliance with algorithm-based ‘Action Plans’. Course content  
93 addresses common gaps that lead to some of the highest sources of global maternal[5,6] and  
94 newborn mortality. [1,2]

95 The competence of participants in these courses are frequently assessed using Objective  
96 Structured Clinical Exams (OSCEs). A number of studies in a variety of LMIC settings have  
97 demonstrated improvements in provider competency managing relevant obstetric and neonatal  
98 cases post training.[6-16] However, few of these studies provide details of assessor training, or  
99 the reliability of the OSCE assessments. [10,15,16] Furthermore, only one study used in-country  
100 OSCE raters;[15] others have relied on external (from outside the country of study) development  
101 and academic partners serving in rater roles.[9,16] Training of raters to serve as OSCEs assessors  
102 is widely embraced in HIC,[17-25] but rater training has not been well described in LMICs.  
103 Reisman and colleagues refer to standardized OSCE training but do not report details.[15] Formal  
104 pre-OSCE training for assessors aims to minimise sources of measurement error, [17-25]  
105 increasing confidence that a participant’s OSCE score truly reflects their competence. With OSCE

1  
2  
3 106 administration, sources of error can arise from the OSCE structure and/or rater  
4  
5 107 objectivity.[17,19,22,25] Facilitator materials for HBB, ECEB and BAB courses provide clear  
6  
7 108 guidelines to minimise measurement error with the OSCE administration. For example, Jhpeigo  
8  
9 109 provides information on quality assessment[3] for their HMS training series, but there are no  
10  
11 110 guidelines for training OSCE raters or evaluating rater agreement. The purpose of our study was  
12  
13 111 to describe a simulation-based OSCE rater training curriculum and assessment of subsequent  
14  
15 112 levels of rater agreement with administration of OSCEs in rural Tanzania using locally trained  
16  
17 113 healthcare providers as raters.  
18  
19  
20

## 21 114 **METHOD**

22  
23  
24 115 This study was embedded within a Simulation Enhanced Maternal Newborn Health training  
25  
26 116 workshop, conducted as part of an ongoing rural education program. The study was approved by  
27  
28 117 Catholic University of Health and Allied Sciences Ethics Board (#CREC/070/2015), the Tanzania  
29  
30 118 National Institute for Medical Research (NIMR) (#MR/53/100/525), and University of Calgary  
31  
32 119 Science and Ethics Board (#REB15-1919).  
33  
34

## 35 120 **Patient and Public Involvement**

36  
37  
38 121 Patients were not involved in this study.  
39

## 40 122 **Setting**

41  
42 123 The study was conducted in Kwimba District located in Mwanza Region, Tanzania over three  
43  
44 124 days; two days in April 2018 and one day in May 2018.  
45  
46

## 47 125 **Participants**

48  
49 126 Raters were recruited from clinical staff practising in the rural health facilities in the district where  
50  
51 127 training was to occur. Selection was based on their demonstrated proficiency in previous Newborn  
52  
53 128 Maternal training workshops conducted in the previous year. All trainees were clinically active in  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 129 their health facility settings. All selected participants provided informed consent to be involved in  
4  
5 130 the study. Rater characteristics and Rater OSCE scores for each OSCE scenario were collated  
6  
7  
8 131 under a master tracking number to ensure rater anonymity. Following three days of rater training,  
9  
10 132 participants were involved as raters for OSCE evaluations to assess workshop learners pre and post  
11  
12 133 training, at 6 and at 12 months. The rater training curriculum was led by a team comprised of  
13  
14 134 clinician researchers from Catholic University of Health and Allied Sciences (CUHAS) and  
15  
16 135 University of Calgary.

### 19 136 **Design**

21 137 This study used a descriptive study design (Figure 1). Raters attended rater training prior to any  
22  
23 138 formal scoring of workshop participants. Categorical levels of proficiency (poor, acceptable and  
24  
25 139 excellent) (decided a priori) were role modelled by clinician research team members for each  
26  
27 140 OSCE each day to create a mock scoring context. All six raters observed and scored the exact  
28  
29 141 same scenario at the same time, making judgements about observed behaviors independent of  
30  
31 142 discussion with each other. Scores were collected and then reviewed with the raters; areas of  
32  
33 143 disagreement were explored, using an inquiry approach for debriefing. Zabar's review criteria  
34  
35 144 and mitigation strategies was used as the framework for both the reviews and refining  
36  
37 145 methodology. . The research team lead (content expert) gave direct feedback Categorical levels  
38  
39 146 of proficiency that challenged rater agreement were repeated. Checklists were collected and  
40  
41 147 collated on an MS Excel spreadsheet on a research dedicated computer. Field notes were used to  
42  
43 148 track challenges. SPSS version 26 was used to analyse rater data. Descriptive statistics were  
44  
45 149 used to provide information about mock scoring and rater's abilities to identify the three  
46  
47 150 categorical levels of proficiency. All raw scores indicating excellent levels of proficiency (Table  
48  
49 151 2) were also analyzed as acceptable (Table 3) to align with training program guidelines; two  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

152 categories of proficiency. Fleiss' Kappa with standard error was calculated to provide  
153 information about the level of rater agreement.<sup>27</sup> Kappa values of <0.20, 0.21-0.40, 0.41-0.60  
154 and 0.61-0.80 and 0.81 to 1.00 are considered poor, fair, moderate, good, and very good  
155 respectively.[27]

## 156 **Evaluation Tools**

157 The OSCEs used were drawn from training program materials. [1-4] There were 24 pass/fail items  
158 on the HBB OSCE, 15 items on the ECEB OSCE and 14 items on the BAB OSCE. All raters were  
159 familiar with the OSCE checklists and relevant training course content as they had recently  
160 participated in the same courses themselves as learners. Poor proficiency, often referred to as 'red'  
161 in reported studies was identified by a score of <71%; 0-17, 0-10, and 0-9 on the HBB, ECEB and  
162 BAB OSCE, respectively. Learner scores >70% identified an 'acceptable' level of proficiency or  
163 'green' in reported studies; >17, >10, & >9 on HBB, ECEB and BAB respectively.<sup>1-4</sup> The Research  
164 team added a third category, a candidate's score of >22, >13 and >12 identified excellent  
165 proficiency for HBB, ECEB and BAB OSCEs, respectively. To standardize the proficiency level  
166 deemed to be acceptable in a scenario, a priori the researchers used the clinical consequences of  
167 an action to inform the scoring, which was then used to plan the actions role played in the scenario.

## 168 **The Rater Curriculum**

169 The conceptual framework (Figure 2) and Zabar's review criteria[28] provides details about  
170 elements of the curriculum and the iterative nature of the training process. Three physical OSCE  
171 stations were set up to facilitate learner transition between each testing station. Checklists were  
172 reviewed prior to scoring practice Day 1 of training to ensure raters were familiar with OSCE  
173 items and how to use the checklist in scoring. Raters observed a scenario, with a predetermined  
174 level of proficiency. Training of raters occurred in the score review, with faculty leading



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

## 179 RESULTS

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

184 **Table 1.** Kappa values

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

186  
 187 The time needed for each OSCE station with score review was longer for average proficiency  
 188 levels (30-40 minutes) when compared to 'excellent' and 'poor' proficiency levels (15-20  
 189 minutes). Fleiss' Kappa values (Table 1) showed that there was a moderate level of rater  
 190 agreement in identifying 'poor' and 'acceptable' proficiency across all OSCEs ( $\kappa=0.51$ ). Kappa  
 191 values improved over the three days moving from 'fair' to 'moderate' for the HBB OSCE and

1  
2  
3 192 'moderate' to 'good' for the ECEB OSCE. The kappa value for BAB was 'moderate' Day 1 but  
4  
5 193 decreased to 'fair' Day 2 and Day 3. Information about rater abilities to correctly identify  
6  
7 194 proficiency levels is described in Table 2 and 3.  
8  
9  
10 195

---

Confidential: For Review Only

196

**Table 2.** 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%)

197

**Table 3.** Proficiency level identification: (average and excellent categories combined) for training program 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%)

198

199 Raters were more accurate in identifying ‘poor’ and ‘excellent’ compared to average, and often  
 200 identified excellent proficiency level scenarios as average. Raters identified average proficiency  
 201 approximately 50% of the time (Table 2 and 3). Information detailing challenges from field notes  
 202 are presented in Table 4.

**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 multi-step OSCE items	Item 1. Prepares area for delivery  Item 2. Equipment preparation Item 3. Hand washing Item 5. Removes wet clothes Item 24. Communication and teaching	Item 7: Improves thermal care  Item 8: Identifying danger signs Advanced care classification Item 10. Medication calculation and administration	Item 7. Controlled cord traction counter pressure Item 12. Determining Postpartum hemorrhage
OSCE English Words		Hypothermia	Hypertension
Actions without verbalizing		Warming baby	

203

## 204 DISCUSSION

205 This study describes an OSCE rater training curriculum and presents evaluation of the  
 206 curriculum showing levels of rater agreement for HBB, ECEB and BAB training courses in an  
 207 LMIC. Quality rater training and subsequent reliability analysis is especially important in LMIC

208 context because of the limited quality assurance monitoring patient safety in the system and  
209 resources.[29-31] Our results suggest that the moderate levels of rater agreement, coupled by  
210 notable challenges in discriminating ‘acceptable’ performance, exposes a potential for either  
211 overestimating or underestimating competence. This has consequences for the individual, the  
212 training program, and the system. The challenge incurred in discriminating between borderline  
213 performance is not isolated to an LMIC context but reported universally.[32-34] With  
214 overestimation of competence, training programs may have passed clinicians who may need  
215 more training to provide safe care on the frontline. The problems of accurate discrimination of  
216 competency also affect resource utilization: with underestimation of competence, training  
217 programs may be directing the limited resources to clinicians who do not need extra training.  
218 Further, frontline staff frequently work short staffed when someone is away at training, so that  
219 unnecessary remediation training may exacerbate staff overload. [26,29-31]

220 In the majority of HBB, ECEB and BAB training program reports, validation of  
221 improved care-giver competency is determined by comparing pre and post training OSCE scores.  
222 Our results suggest that the existing reports describing a moderate IRR may be misleading  
223 without further validation of the accuracy of rater discernment of acceptable proficiency.  
224 [10.15.16] Our raters achieved moderate rater agreement yet discernment of acceptable  
225 proficiency, which is the pass criterion in these training programs, was approximately 50%.  
226 Based on our findings we would suggest including both measures of validation. Considering  
227 contexts with limited resources, it may be helpful to implement a further strategy such a global  
228 rating scale, which is common practice in HICs [17-19,22-25] to provide another method of  
229 validation of participant competence. [27,28] A global rating scale allows the rater to evaluate  
230 how well a learner performs on a scale of 1 to 5, with 5 reflecting the highest level of

1  
2  
3 231 competence.[28] More than one method of validation creates more certainty that results are an  
4  
5 232 accurate reflection of participant competence and/or training program efficacy [27]. With the  
6  
7  
8 233 continued high reports of maternal and neonatal mortality, it is important to be confident that  
9  
10 234 these training programs are accurate in identifying and supporting clinicians who may not be  
11  
12 235 providing safe care on the frontline.  
13

14 236  
15 237 The guidelines for OSCE rater training used in this study were based on  
16  
17  
18 238 recommendations from HIC rater training experiences; these are challenging to implement in an  
19  
20 239 LMIC context. Globally, good practice is for OSCE raters to have relevant content expertise, be  
21  
22  
23 240 well orientated to the OSCE checklist and use a validated rating scale.[22-25] Although we  
24  
25 241 strived for this, we had a limited pool of potential raters; this may have affected the challenges  
26  
27 242 we noted in rater perceptions of the expected practice standard. Raters were recruited by  
28  
29 243 clinician researchers based on recollections of which previous participants from recent HBB,  
30  
31 244 ECEB, and BAB trainings had performed well; no objective strategy was employed in their  
32  
33 245 selection. This was the reason in country faculty inserted a third categorical level of proficiency;  
34  
35 246 excellent. They wanted an objective strategy to identify content experts as the future raters for  
36  
37  
38 247 such training programs. A quality rater training curriculum includes standardized mock  
39  
40  
41 248 scenarios where raters practise with a variety of expected learner proficiency levels demonstrated  
42  
43 249 and practice scored. In our study, this was one of the greatest challenges. Research Clinicians  
44  
45 250 role playing scenarios Day 1 were challenged in demonstrating poor proficiency. In discussion,  
46  
47 251 they shared they didn't want participants to think they were not experts in the field. The  
48  
49  
50 252 inclusion of scripted and video capture of proficiency levels may lessen this tension and  
51  
52 253 inconsistency in role play. Despite this, the level of rater agreement improved over the three  
53  
54  
55 254 training days for both HBB and ECEB. The fall-off in rater agreement for BAB Day 3 was  
56  
57  
58  
59  
60

255 unexpected but may be in part related to the timing of these scenarios Day 3; they were the last  
256 role plays of the day and rater fatigue may have played a role. Additionally, the greater number  
257 of differing perceptions of the practice standard (Table 4) may have impacted this finding.

258 A solid rater curriculum incorporates a framework such as Zabar's (Figure 2) to guide  
259 rater feedback; this is especially important in a setting where the concept of rater training is  
260 novel. In our study, Zabar's framework was simple and easy to use as evidenced by a decreased  
261 level of external coaching each day.

262 A study strength was the achievement of a level of rater agreement similar to the few  
263 published training course reports for ECEB and HBB. In our participant group, the 'moderate to  
264 good' kappa for the ECEB OSCE was as reported by Kassick and colleagues in Ghana, the only  
265 other ECEB reported study to include in-country evaluators; a regional and national  
266 evaluator.[10] In the HBB OSCE, our findings demonstrated 'fair to moderate' kappa value  
267 which was similar to the 'fair to good' kappa value reported by Reisman and colleagues in  
268 Tanzania[15] whose raters included two external evaluators and one country based evaluator.  
269 Comparable studies for kappa value results for raters scoring the BAB OSCE module are not  
270 reported. The achievement of comparable IRR to the studies using in country and external  
271 partners provides support for the rater training curriculum, yet the inability to accurately discern  
272 acceptable proficiency (pass criteria) is concerning. To gain further insight into the relationship  
273 between faculty role play and the inability to discern acceptable proficiency, we plan to script the  
274 acceptable proficiency level for each OSCE, coach faculty in the role play, and repeat the  
275 curriculum and analysis.

276 Rater trainees were challenged by OSCE items where scores incorporated multi-steps for  
277 their achievement; this was consistent with experiences described by Seto and colleagues who also

1  
2  
3 278 identified lower rater agreement for HBB OSCE multi-step items.[16] For example, in our study,  
4  
5 279 one HBB OSCE 'item' requires the learner to 'prepare the area for delivery'. To achieve a point  
6  
7  
8 280 and 'pass' this item, the learner must complete all four of: (1) place towels at bedside; (2) place  
9  
10 281 suction at bedside; (3) place a bag and mask at bedside; and (4) place oxytocin at bedside. This  
11  
12 282 'item' created confusion amongst rater trainees; during mock session review, several participants  
13  
14 283 had 'passed' the mock scenario learner on this item despite not having seen all steps yet having  
15  
16 284 observed at least one step. To address this gap, we added sub-item tracking boxes when this  
17  
18  
19 285 challenge was identified Day 1; the use of this strategy warrants further study.  
20  
21

22 286 Our study was limited by lack of formal training and experience in role-playing by  
23  
24 287 simulated learners. Our 'actors' were not professionally trained (but rather research clinicians!)  
25  
26 288 and scenarios and levels were de novo; ideally, with more resources and time, mock scenarios  
27  
28 289 would be formally scripted and/or video-captured to optimize standardization. Additionally, time  
29  
30 290 constraints necessitated working three long days; rater fatigue was likely. This was especially true  
31  
32 291 for one pregnant rater-trainee who participated for the first two days then arrived with newborn in  
33  
34 292 hand on Day 3. Our results may have limitations in generalisability but do provide some context  
35  
36 293 and learning for others interested in developing a rater training curriculum in a low resource  
37  
38 294 setting.  
39  
40  
41  
42 295

## 44 296 **CONCLUSION**

46 297 Our results show that rater training in an LMIC setting is critical for administering OSCE based  
47  
48 298 learner assessments especially since the raters in this study had a hard time identifying average  
49  
50 299 performance. Clinician everywhere need ongoing training, but to optimize learning and then  
51  
52 300 translate this to improved outcomes for mothers and babies, this training must be informed by truly  
53  
54  
55  
56  
57  
58  
59  
60



1  
2  
3 301 objective evaluations. Our study shows in rural, Tanzania, training of in-country raters is possible  
4  
5 302 and can lead to an IRR which is similar to previous studies. Improved standardization and attention  
6  
7 303 to the relationships between IRR and the accurate discernment of participant performance would  
8  
9 304 provide insight into needed modifications, which in turn may lead to greater accuracy in rating  
10  
11 305 competence. More research is warranted. Global training programs, including HBB, ECEB and  
12  
13 306 the BAB need to be confident that OSCE scores truly reflect learner ability, to identify and support  
14  
15 307 those needing further skill practice. Significant global investments have been made towards  
16  
17 308 maternal newborn health provider training; participants need to leave workshop venues equipped  
18  
19 309 with the skills to save mother and newborn lives. We hope this experience encourages program  
20  
21 310 developers nationally and internationally to scale up in-country rater training. For LMIC  
22  
23 311 simulation-based training programs to be sustainable, all countries and regions should have their  
24  
25 312 own trained OSCE raters.  
26  
27  
28  
29  
30  
31 313

## 315 Funding

316 “This work was supported by a grant from the Innovating for Maternal and Child Health in Africa  
317 (IMCHA) initiative- a partnership of Global Affairs Canada (GAC), the Canadian Institutes of Health  
318 Research (CIHR) and Canada’s International Development Research Centre (IDRC), **Grant number**  
319 **108024-001** under Mama na MToto programme in rural Tanzania.

320  
321 Study sponsors had no involvement in study design, collection and analysis of data, interpretation or  
322 writing of this manuscript.

## 324 Acknowledgements

325 Healthcare workers from Misungwi District who served as raters for this training program.

326

327

328

329

330

What is already known:

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

336

What this study adds:

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

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

## 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, 2<sup>nd</sup> edition. 2015. Available from <https://www.aap.org/en-us/advocacy-and-policy/aap-health-initiatives/helping-babies-survive/Pages/Helping-Babies-Breathe.aspx> (Accessed 19 March 2018)
3. Jhpeigo Helping Mothers Survive Training Skills for Health Care Providers, Third Edition: Reference Manual. Editors: Bluestone J, Fowler R, Johnson P, Smith J. Published by Jhpeigo Corporation, USA. 2010. Available: [http://resources.jhpiego.org/system/files/resources/trainingskills\\_manual\\_0.pdf](http://resources.jhpiego.org/system/files/resources/trainingskills_manual_0.pdf).
4. Jhpeigo. Helping Mothers Survive Bleeding After Birth Training Package. 2016 Available from <http://reprolineplus.org/resources/HMS>. (Accessed 19 March 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. *International Journal of Gynecology & Obstetrics* 2014;126(3):286-90.
7. Beena D, Kamath-Rayne BD, Thukral A, et al (2018). Helping Babies Breathe, Second Edition: A Model for Strengthening Educational Programs to Increase Global Newborn Survival. *Global Health: Science and Practice*;2018; 6(3): 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;93:287–295.
9. Brucker M. Management of the third stage of labor: an evidence-based approach. *J Midwif Women’s Health*. 2001;46:381-92
10. Kassick M, Chinbuah M, Serpa M, et al. Evaluating a novel neonatal-care assessment tool among trained delivery attendants in a resource-limited setting. *International Journal of Gynecology and Obstetrics* 2018;135(3):285-89.
11. Alwy F, Pembe AB, Hirose A, et al. Effect of the competency based Helping Mothers Survive Bleeding after Birth (HMS BAB) training on maternal morbidity: A cluster randomized trial in 20 districts in Tanzania. *British Medical Journal Global Health* 2019;4(2):e001214.
12. Bishanga DR, Charles J, Tibaijiuka G, et al. Improvement in the active management of the third stage of labor for 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 Practice & Research Clinical Obstetrics and Gynaecology*;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(5):300–08.

- 1  
2  
3 396  
4 397  
5 398  
6  
7 399  
8 400  
9 401  
10 402  
11 403  
12 404  
13  
14 405  
15 406  
16 407  
17 408  
18 409  
19 410  
20 411  
21  
22 412  
23 413  
24 414  
25 415  
26 416  
27 417  
28 418  
29  
30 419  
31 420  
32 421  
33 422  
34 423  
35 424  
36 425  
37 426  
38 427  
39 428  
40 429  
41 430  
42 431  
43 432  
44 433  
45 434  
46 435  
47 436  
48 437  
49 438  
50 439  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60
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. *International Journal of Gynecology and Obstetrics* 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. *Perspectives on Medical Education* 2015; 4(5):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(9):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. Objective structured clinical examinations. Oxford University Press, Oxford UK. 2013.
  20. Van Der Vleuten CPM, Van Luyk SJ, Van Ballegooijen AMJ, et al. Training and experience of examiners. *Med Educ* 1989;23:290–96.
  21. Harden, RM. "Revisiting 'Assessment of clinical competence using an objective structured clinical examination (OSCE)'" *Med Educ* 2016;50(4): 376–79.
  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(4): 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 Medical Education* 2017;17(71): 1-7.
  24. Pugh Vijay John Daniels & Debra. "Twelve tips for developing an OSCE that measures what you want." *Medical Teacher* 2018;40(12):1208-13.
  25. Preusche I, Schmidts M, Wagner-Menghin M. 2012. "Twelve tips for designing and implementing a structured rater training in OSCEs." *Medical Teacher* 2012;34(5):368-372.
  26. 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." 1-154. Available: <https://dc.sourceafrica.net/documents/118198-Tanzania-Health-Sector-Strategic-Plan-July-2015.html> (Accessed 19 March 2017).
  27. Fleiss JL, Cohen J. "The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability." *Educational and Psychological Measurement* 1973;33: 613-619.
  28. Zabar S, Krajic Kacher E, Kalet A, et al. 2013. *Objective Structured Clinical Exams- 10 steps to planning and implementing OSCE's and other standardized patient exercises*. Edited by Kachur E, Hanley K. Zabar S. New York: Springer.
  29. World Health Organization, OECD, and International Bank for Reconstruction and Development/The World Bank, 2018. 2018. "Delivering quality health services. A global imperative for universal health coverage."
  30. 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](https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(18)30386-3/fulltext) (Accessed 6 November 2018).

- 1  
2  
3 440 31. Rowe AK, Labadie G, Jackson D, et al. 2018. "Improving health worker performance an  
4 441 ongoing challenge for meeting the sustainable development goals." *BMJ* 2018 (362):  
5 442 k2813.  
6  
7 443 32. Fuller R, Homer M, Pell G, et al. (2017) "Managing extremes of assessor judgment within  
8 444 the OSCE." *Medical Teacher* 2017;39(1):58-66.  
9 445 33. Petrusa ER. (2002). 'Clinical Performance Assessments' in Norman G, van der Vleuten C,  
10 446 Newble D (eds) *International Handbook of Reserach in Medical Education* Boston,  
11 447 Kluwer Academic Publishers.673-709.  
12 448 34. Reid K, Smallwood D, Collins M, et al. "Taking OSCE examiner training on the road:  
13 449 reaching the masses." *Medical Education Online* 2016;21(1).  
14  
15 450  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Confidential: For Review Only

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

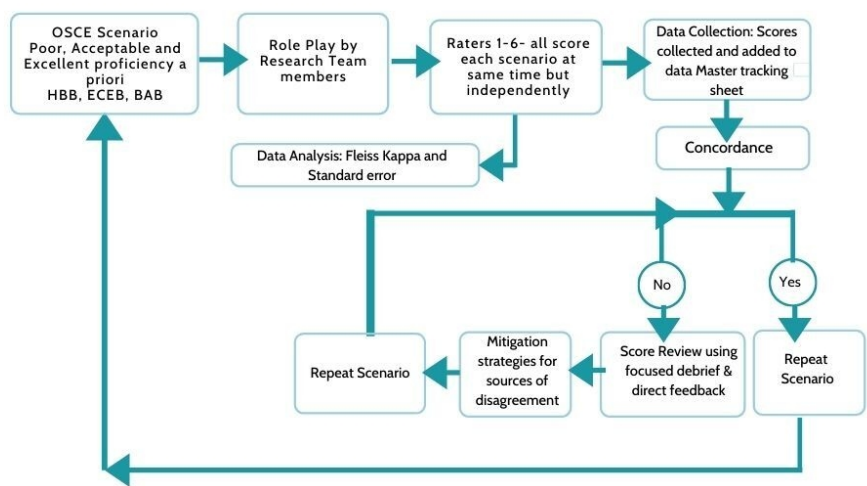


Figure 1. This figure provides a visual of the research design we used in the study each day. All six raters scored all 42 of the role played scenarios with proficiency determined a priori. Raters participated in 42 debrief sessions over the three days.

43x32mm (600 x 600 DPI)

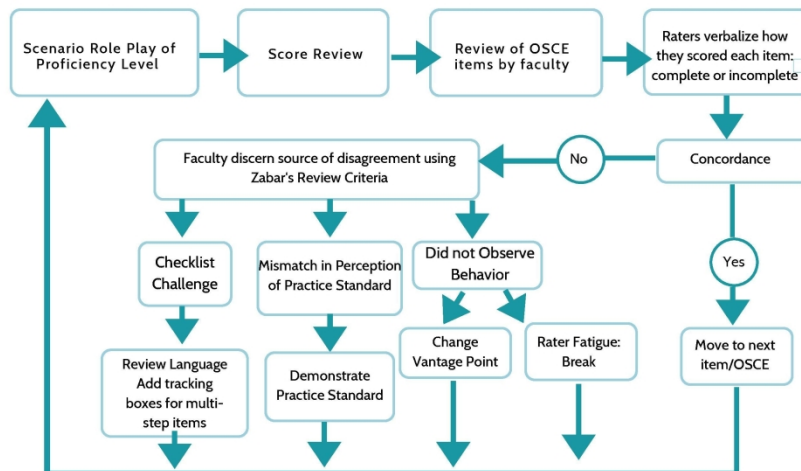


Figure 2. This figure provides a visual of the Conceptual framework used to improve the level of rater agreement.

135x101mm (600 x 600 DPI)

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60