Supplemental material 17. Results Chapter 2 publication

The application of Item Response Theory to revolutionise analysis of clinical competence of health workers: a cross-sectional vignette-based EmONC competence test of Burundian delivery care professionals

Desire Habonimana[1,2], Attakrit Leckcivilize[2], Catia Nicodemo[3], Jean Baptiste Nzorironkankuze[4], Ananie Ndacayisaba[5], Aristide Bishinga[6], Jeanine Ndayisenga[7], Eugenie Siga Diane Niane[8], Sylvestre Bazikamwe[9], Pontien Ndabashinze†[10], Mike English[2]

[1]Centre de Recherche Universitaire en Santé (CURSA), Department of Community Medicine, Faculty of Medicine, University of Burundi, Bujumbura, Burundi.

[2]Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK

[3]Nuffield Department of Primary Care and Health Science, University of Oxford, Oxford, UK and University of Verona, Italy

[4]Ministry of Health, Republic of Burundi, Bujumbura, Burundi

[5]Reproductive, Maternal, Child, Adolescent, and Newborn Health programme, Ministry of Health, Republic of Burundi, Bujumbura, Burundi

[6]Maternal and Child Health programme, Japan International Cooperation Agency in Burundi, Bujumbura, Burundi

[7]The East African Community Centre of Excellence in Public Heath Training, Department of Clinical Sciences, National Institute of Public Health and Burundian Association of Neonatology, Bujumbura, Burundi

[8]Reproductive, Maternal and Neonatal Health Department, World Health Organisation, Bujumbura, Burundi

[9]Department of Gynaecology and Obstetrics and Burundian Association of Gynaecology and Obstetrics, University of Burundi, Bujumbura, Burundi

[10]Department of Paediatrics and Neonatology and Burundian Association of Neonatology, University of Burundi, Bujumbura, Burundi

Corresponding author: Desire Habonimana | Tel: +257 76 620 326 | E-mail: desire.habonimana@ub.edu.bi

Abstract

Background

Competence is a latent trait difficult to comprehend with frequentist statistics. Using delivery care professionals from Burundian emergency obstetric and newborn care (EmONC) facilities, we demonstrate how item response theory (IRT) provides a deeper analysis of clinical competence of health workers and enables the design and validation of competence test instruments tailored to particular categories of health professionals.

Methods

Based on EmONC clinical guidelines, the Burundian EmONC training curriculum, and contributions from 23 maternal and newborn health (MNH) experts, we designed five EmONC clinical vignettes representing the most common maternal and newborn complications and examined 311 delivery care professionals across all 112 Burundian EmONC facilities. We compared summated scores and one-way ANOVA results with IRT findings, documenting analytical merits of IRT features particularly the item and total characteristic curves (ICC and TCC) used to apprehend the test difficulty and predict expected scores assuming an infinite repeat of test administration. Furthermore, we estimated the latent competence using empirical Bayes means and tested the test instrument validity using the test information function (TIF).

Results

Summated statistics and the IRT model produced comparable overall results but the latter offered more nuanced insights. Item Characteristic Curve (ICC) analysis identified the most challenging test components. Analysis of Variance (ANOVA), Test Characteristic Curve (TCC) findings, and empirical Bayes means estimates collectively demonstrated that higher qualifications positively impact EmONC competence. Based on ANOVA results, nurses scored 44% on the EmONC competence test, midwives 56%, and doctors 70% while TCC indicated that random samples of nurses, midwives, and doctors in Burundian EmONC facilities would achieve 45%, 57%, and 62% respectively. Correspondingly, the empirical Bayes means confirmed that doctors have EmONC competence $\theta=1.13$ (SD=0.88), midwives $\theta=0.40$ (SD=0.84), and nurses $\theta=-0.12$ (SD=0.95) on $\theta\pm4$ scale. Finally, the test instrument provided precise estimates for professionals with $-0.5 < \theta < 0.5$ and was therefore suited for midwives and nurses.

Conclusions

Although ANOVA and IRT findings corroborated one another, this methodological paper demonstrated that IRT allows us to discriminate participants based on question difficulty levels and to predict underlying competence which frequentist statistical methods cannot accommodate.

Keywords

Item response theory, EmONC competence, delivery care providers, Burundi

Background

Competent delivery care professionals in Burundian Emergency Obstetric and Newborn Care (EmONC) facilities are crucial to saving lives. It is estimated that 340 Burundian women and 2,300 newborns die for each 100,000 deliveries, mainly from obstetric and newborn complications that could be averted if quality obstetric care is delivered [1-3]. The Sustainable Development Goals (SDGs) agenda highlights the need for sufficient skilled delivery care professionals to deliver on the maternal and neonatal agenda [4, 5], and mounting evidence underscores that quality care is only possible with competent health workers capable to correctly perform salient clinical tasks [6, 7]. Unfortunately, research has found that a significant proportion of qualified health workers specifically in low- and middle-income countries (LMICs) have insufficient skills to offer correct care for most maternal and newborn complications [7-10]. For countries like Burundi which aspires to achieve about 60% and 50% reductions in maternal and newborn mortality rates respectively by 2030, understanding health worker competence in EmONC maternities is helpful to inform actions and policy change.

There are varied approaches to assessing competence, because it is a latent trait it is difficult to apprehend and for its expression often depends on an enabling environment so direct observation during actual clinical task performance would be optimal [11]. However, such observation is very difficult operationally and the most commonly used methods to assess competence in clinical medicine are vignettes, written tests, computer-delivered tests, performance records, physical models, and task samples [6, 12, 13]. Of these methods, vignettes are believed to be less tedious and inexpensive particularly in LMICs where routine chart documentation is often incomplete or absent [14].

Statistical analyses of competence based on vignettes have typically relied on classical test theories such as summative scores and principal component and factor analyses [8, 15-18], but these frequentist methods do not accommodate important hypotheses underpinning this form of assessment. First, competence test components generally differ substantially in their degree of difficulty perhaps particularly with obstetric vignettes in that some routine clinical practices are generally intuitive (e.g., measuring blood pressure) while other tasks require advanced clinical expertise (e.g., instrumental childbirth and neonatal resuscitation) [19, 20]. Second, assuming that they don't recall previous answers, we would expect examinees to fare differently if the competence test was administered repeatedly over a relatively short period of time, and these scores would cluster themselves around some average value corresponding to an individual's 'true' score [19]. Therefore, simpler statistical analyses particularly those using equal weighting fail to discriminate between easy and difficult test components. Further, these methods simply rely on summary scores to estimate competence ignoring that the latter is a latent feature different from obtained scores.

Item Response Theory (IRT), a psychometric competence analysis technique which emanates from the sectors of education and psychology [21], offers a means to address these issues [18, 22-24] while allowing a simultaneous computation of the predicted true scores along with the latent competence [11, 25]. Nonetheless, despite its documented advantages over summary statistics, this form of analysis has rarely been applied to examining competence of health workers and those that used it did not demonstrate methodologically how it can be done using applied statistical packages (e.g., STATA) [11, 26, 27]. This paper demonstrates the application and rewards of IRT in examining competence of health workers using delivery care providers in Burundian EmONC facilities. Specifically, we compare summated and one-way ANOVA scores with IRT findings and use IRT attributes to isolate the most difficult questions and test components, detect how easy and difficult test sections affect performance of different delivery care professionals, and predict the expected scores and the underlying competence assuming an infinite number of test repeats. Finally, we demonstrate how IRT enables us to design new or update existing competence test instruments tailoring them to particular categories of health professionals.

Methods

Study design and setting

We conducted a cross-sectional study to examine EmONC competence among nurses, midwives, and medical doctors working in maternity across all EmONC facilities in Burundi (n=112). These facilities include 53 hospitals and 59 primary health facilities designated as comprehensive and basic EmONC facilities (CEmONC and BEmONC respectively) and comprise 82 (73%) public and 30 (27%) religious and private facilities. In 2021, there were 51 doctors in 32/53 CEmONC facilities, 103 midwives across 49/53 CEmONC and 5/59 BEmONC facilities, and 165 EmONC nurses in 34/53 CEmONC and 51/59 BEmONC facilities specifically posted to maternity and neonatology units. In the same year, the CEmONC facilities performed between 331 and 3737 (median 1622) deliveries and handled from 149 to 2385 (median 876) maternal complications and 29 to 816 (median 247) newborn complications and the BEmONC facilities performed 493 to 2431 (median 1365) deliveries and managed from 77 to 1748 (median 805) maternal complications and 15 to 670 (median 215) newborn complications [28].

Study participants and data collection

In Burundi, all qualified delivery care professionals namely nurses, midwives, and doctors should receive the standard EmONC training to upgrade their clinical skillset linked to the management the most maternal and newborn complications except emergency surgery (i.e., hysterectomy and caesarean section) which requires the competence of medical doctors. It is therefore expected that doctors, midwives, and nurses in these EmONC maternities are familiar with care for routine childbirth and for various emergency complications excluding surgery [28]. Therefore, the EmONC competence test excluded obstetric surgery but targeted all delivery care providers irrespective of their qualifications. This competence test was administered by 54 examiners grouped into 18 teams; one team per province of Burundi. Those administering the test were medical doctors who received standard training and who demonstrated consistent comprehension of the administration of questions during pretesting. Each team of examiners was supervised by an experienced MNH expert (e.g., academic and members of the study stakeholder committee additional file 1) and together visited all EmONC facilities in their assigned province over a period of five days in July 2022; examining between two to four delivery care professionals in each facility. Engagement of stakeholders and the process of study protocol and tools development and validation have been described in our previous report [29].

Sample size calculation

Drawing on evidence from other African and Asian low-income contexts [7, 9, 10, 30] and using the cross-sectional sample size estimation method for prevalence $p = 0.25$ (i.e., we would expect one in four Burundian delivery care professionals to correctly answer the whole EmONC competence test), $z = 1.96$ for $\alpha = 0.05$, and $\varepsilon = 0.05$; we estimated a minimum sample of 289 delivery care professionals was needed. Based on the average workforce capacity of Burundian facilities, we expected 118 delivery care professionals in BEmONC facilities and 212 professionals in CEmONC facilities; thus our sample represented approximately 88% (289/330) of the expected universe of possible delivery care professionals in these designated facilities. Our sample does not include obstetricians, paediatricians, and neonatologists who only exist in tertiary and academic hospitals in Bujumbura and do not provide routine maternity and neonatal care.

EmONC clinical vignettes construction

We designed five EmONC clinical vignettes representing the most common maternal and newborn complications in Burundi (Table 1) based on the guide for health workers on the management of complications in pregnancy and childbirth jointly developed by WHO, the United Nations Population Fund (UNFPA), and the United Nations Children's Fund (UNICEF) [31]. This guide was the basis of the nationally deployed Burundian EmONC training curriculum and the EmONC provider skillset requirements. These vignettes are a battery of questions for hypothetical cases and were designed in collaboration with maternal and newborn health (MNH) clinicians and experts namely researchers, implementers, and professional associations. Examinees were provided with each hypothetical clinical case individually and asked without prompting to proceed with care as they would under actual circumstances. Using vignette 2 as an illustration, the examiner informed an examinee that a pregnant woman presents with symptoms of severe malaria and asked the immediate actions to undertake and the danger signs to monitor. In this case, the latter actions and danger signs represent our $response\ items$ and are fully explained during EmONC training and commonly supplemented in maternity wards as clinical guidelines. All vignettes used a similar approach and the examiner recorded 1 if examinee i correctly provides $response\ item\ j$, and 0 if the response is wrong or not given.

Vignette cases and corresponding sub-questions	Response items (n)
Vignette 1. Labour monitoring (n=3)	**20 response items**
What are the key elements to regularly monitor during childbirth?	10 response items
What are the indications and key components of the active management of the third stage of labour?	5 response items
What are the safety measures to undertake for preventing septic childbirth?	5 response items
Vignette 2. Management of malaria during pregnancy (n=2)	**13 response items**
What are the immediate actions to take when a pregnant woman presents severe malaria symptoms?	5 response items
What are the danger signs of severe malaria during pregnancy?	8 response items
Vignette 3. Management of postpartum haemorrhage (n=5)	**41 response items**
What are the danger signs to monitor in case of abundant maternal bleeding?	7 response items
What are the important and systematic actions to do in case of abundant maternal bleeding?	9 response items
What are the key and systematic actions to do in case of a retained placenta?	10 response items
What are the signs to monitor when a mother develops malaise within 48 hours after delivery?	6 response items
What are the key actions to undertake following an incomplete or an unsafe abortion?	9 response items
Vignette 4. Management of severe preeclampsia and eclampsia (n=6)	**24 response items**
What do you suspect when a mother presents severe headaches and abdominal pain?	4 response items
What are the positive signs to check to confirm suspected severe preeclampsia?	5 response items
What are the essential monitoring elements in case of severe preeclampsia?	6 response items
What are the important and systematic actions to do after confirmation of severe preeclampsia?	4 response items
What are the key signs to assess before a repeat administration of magnesium sulfate (MgSO4)?	3 response items
What are the key actions to do in case of a cardiopulmonary arrest after the administration of MgSO4?	2 response items
Vignette 5. Immediate care for the newborn (n=3)	**17 response items**
What are the systematic elements of newborn care immediately after birth?	8 response items
What are the essential and systematic steps of birth asphyxia treatment?	6 response items
What is the systematic approach of PMTCT?	3 response items
Total number of questions (n=19)	**115 response items**

Table 1 shows the number of questions and response items in each clinical vignette. In total, the whole EmONC competence test comprised five vignettes and 19 questions with a total 115 possible response items. The third vignette concerned the management of postpartum haemorrhage and accounted for 35% of response items (n=41).

Data analysis

Descriptive statistics

We began data analysis with simple explorations, determining proportions of correct responses on each item and identifying items and vignettes that are most challenging. Next, we generated the total raw scores out of 115 (the maximum possible score with equal item weighting) for each participant and summarised them by category of delivery care professionals and further by facility type (BEmONC versus CEmONC) focusing on midwives and nurses, as doctors are generally only posted to CEmONC facilities. We explored the inter-item correlations within and across vignettes and examined the test reliability using Cronbach's alpha. Group summaries of raw scores were compared using one-way ANOVA to examine the mean score differences between nurses, midwives, and doctors for each vignette and on the whole set of vignettes. We aggregated and summarised crude scores by health district and by province, principally to explore whether specific provinces had lower mean scores to inform further EmONC training needs using a forest plot to illustrate the findings.

Item Response Theory analytical concepts

Each examinee responding to a test item possesses some level of underlying ability (i.e., latent prior competence) that determines the probability of a correct answer to that particular item (Equation 1) [32]. Conceptually, IRT updates the latter prior competence with actual test scores to predict the posterior competence and further uses the empirical Bayes means estimates and the standard errors to examine the validity of the overall test instrument [32]. This method enables us to distinguish the levels of test difficulty for each item and to predict true scores that an examinee with a given level of prior competence would achieve if the test were to be administered an infinite number of times [33]. Thus, IRT is particularly helpful to determine whether the test instrument is powerful for examining competence across all test takers or whether it is more robust to capture this competence in particular categories of these examinees (e.g., such as nurses) [34]. For brevity, the terms ability and competence are employed interchangeably, true score denotes the predicted success rate assuming an infinite repeat of test administration, and true competence designates posterior competence which, together with the item difficulty, are continuous metrics that can range from $-\infty$ to $+\infty$.

$$TS_i = \sum_{j=1}^{J} P_j(\theta_i) \qquad \text{Equation 1.}$$

where

TS_i	is the true score for examinees with ability level θ_i
j	corresponds to an item; J is total number of items
$P_j(\theta_i)$	is the probability of success to item j which depend to ability level θ_i.

Test difficulty and predicted test scores

Response items were dichotomous taking value 1 if respondent i correctly gives $item\ response\ j$, and 0 otherwise; and the total raw score for respondent i ranged between 0 and 115. We denoted b_j to represent the difficulty of $response\ item\ j$ and θ to denote the examinee's underlying EmONC competence. Therefore, each response to an item represents an encounter between examinee of ability θ and an item of difficulty b_j such that professionals with higher θ than b_j have the likelihood of scoring 1 [22, 35]. This relationship between θ and b_j further implies that if $\theta < b_j$, the probability of success to $item\ response\ j$ tends to 0 and becomes 0.5 if they are equal (Equation 2). We used a one-parameter logistic model (1PL) and the item characteristic curves (ICC) to determine item difficulties, describing the probability that respondent i succeeds on each response item. The ICC uses three parameters of the logistic functional form of the IRT model; the $guessing$ parameter which represents the probability that a provider with the lowest possible competence will correctly respond to item j, the $discrimination$ parameter which corresponds to the maximum slope of the curve, and the $difficulty$ parameter corresponding to the competence level θ at which the probability of responding correctly to item j equals 0.5. Because we assume a zero mean for θ, response item j is believed to be relatively easy if its difficulty estimate is negative and relatively hard if its difficulty estimate is positive. For comparison purposes, we reported each item difficulty level along with the proportion of examinees who succeeded on it.

$$\Pr\left(Y_{ij} = 1 \middle| b_j, \theta_i\right) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)}$$ Equation 2.

where

Y_{ij}	represents the (yet to be observed) outcome for item j from person i
b_j	corresponds to item j difficulty
θ_i	is the ability level of provider i
$(\theta_i - b_i)$	ranges from - ∞ to +∞

Next, we predicted the expected true score or success rate that examinee i with competence θ_i would achieve if the EmONC test were taken repeatedly. To express this more clearly, given that θ_i is a latent trait being examined using the EmONC competence test, we transformed the latter θ_i into the proportion 'true score', corresponding to the proportion of response items an examinee with θ_i is expected to answer correctly where expectation is the average score over repeated administrations of the same competence test. The total characteristic curve (TCC) function of the IRT model plots expected scores on the competence test along the latent trait (θ) continuum. As our EmONC competence test contained 115 binary response items each coded 0 or 1, an examinee i would score between 0 and 115. We used the 95% critical threshold from a standard normal distribution (z = ± 1.96) to predict the expected average (and the minimum and maximum) scores any random sample of nurses, midwives, and medical doctors would achieve on this EmONC competence test.

Finally, we updated the posterior latent competence by predicting θ for each test taker using the empirical Bayes means estimates which combine prior information about the latent competence θ with the likelihood function (actual test scores) to obtain the conditional posterior θ distribution. The posterior θ represents the true competence and generally ranges between —4 and 4 with participants with θ in the vicinity of zero being believed to be averagely competent. We examined the distribution of true competence by category of delivery care professionals and by facility type and further investigated the effects of qualification, previous EmONC training, delivery experience, type of employment, sex, and age of professionals on competence using linear regression. This regression included quadratic terms for age and previous deliveries to examine possible non-linear relationships with competence. Specifically, we ran complete and then saturated linear regressions each containing an additional dichotomous predictor variable coding for providers with and without EmONC training by type of professionals to explore the effects of qualification and EmONC training on true competence. Diagnosis of linear regressions detected heteroskedasticity so we predicted robust estimates with 10,000 bootstraps to improve standard errors [36].

EmONC competence test validity assessment

Although our EmONC competence test drew on clinical EmONC guidelines [31] and the standard EmONC training curriculum and further benefited from contributions of MNH experts, our instrument (i.e., the clinical vignettes) has never been tested and validated before. We were therefore uncertain whether this instrument is accurate in capturing EmONC competence among Burundian doctors, midwives, and nurses at the same time even though these professionals receive the same EmONC training apart from emergency obstetric surgery which concerns medical doctors and was not part of the competence exam. Test validity was diagnosed using the test information function (TIF) of the IRT model; determining the competence range where this instrument provides more information by minimising standard errors. This maximum precision range therefore reveals the types of participants whom the test is particularly suitable for. Indeed, moving away from that precision interval in either direction increases the standard error and the instrument becomes inaccurate; providing less information about θ.

Results

Descriptive statistics and raw data analysis results

We examined a convenient sample of 311 professionals comprising 257 nurses, 40 midwives, and 14 medical doctors who were on duty during our visits. About half of participants came from either type of facility (159/311, 51% in BEmONC facilities and 152/311, 49% in CEmONC facilities), but BEmONC facilities had more nurses (152/257, 59%) while CEmONC facilities had more midwives (34/40, 85%) and the vast majority of doctors (13/14, 93%). Higher qualified professionals were predominantly male with 152/257 (60%) female nurses against 26/40 (65%) male midwives and 13/14 (93%) male doctors. On the whole, one in three examined professionals was below 35 years and 95% of all participants were below 50. We found that 96% (300/311) of professionals had more than 2 years of maternity experience although above half (161/311, 52%) had never received EmONC training before.

Crude test results in additional file 2 show that few delivery care professionals in Burundian EmONC facilities have the needed competence across all clinical vignettes. Respondents performed better in vignettes on labour monitoring (64.3%) and routine neonatal care (58.6%) and worse on those regarding the management of severe preeclampsia and eclampsia (35.6%), fever and malaria (36.5%), and postpartum haemorrhage (45.4%). With respect to specific clinical tasks, only 17% of examinees proved to be competent to manoeuvre the perineum during the active management of the third stage of labour and to identify standard danger signs of severe malaria during pregnancy. Further, this additional file 2 shows that not one of the examined professionals knew the respiratory rate criterion that indicates whether it is safe to repeat doses of magnesium sulphate (MgSO4) during the management of hypertensive disorders and only one in twelve would regularly assess clotting signs to rule out coagulopathy during MgSO4 treatment.

Response items correlated positively within and across clinical vignettes with an overall Cronbach's reliability criterion α = 0.945. Additional file 3 demonstrates that the overall percent mean raw score increases with higher qualification with nurses achieving 44.0% on average, midwives 55.7%, and medical doctors 69.9%. Generally, doctors fared better in performance than midwives and nurses particularly concerning the more severe complications (e.g., hypertensive complications and postpartum haemorrhage) while nurses or midwives demonstrated slightly better or comparable competence to doctors in the management of routine care (e.g., prevention of mother-to-child transmission of HIV). On the whole, we found that all professionals have good clinical competence in labour monitoring (from 61.6% among nurses to 84.3% among medical doctors) and infection control during childbirth (from 59.2% among nurses to 82.9% among doctors). The forest plot of province-level mean competence levels is provided as additional file 4 and suggests that some provinces may have less competent staff but on a background of need for greater EmONC competence and resource provision nationally [28].

Difficulty of clinical vignettes and vignette items

Findings of the item difficulty analysis are summarised in Tables 1 to 5 of additional file 2 and demonstrate that response items covered a wide difficulty spectrum, with the first item being the easiest (b_1 = —3.08) and item 74 the most difficult (b_{74} = 2.58). The first clinical vignette concerned labour monitoring and suggests that averagely competent respondents (i.e., those with θ in the vicinity of zero) would correctly respond to 17 out of 20 items. Similarly, averagely competent professionals would correctly answer 70% of the immediate newborn care items (vignette 5). The hardest EmONC vignettes are the management of fever and malaria (vignette 2), postpartum haemorrhage (vignette 3), and hypertensive disorders (vignette 4). Generally, any delivery care provider would require a high level of competence ($\theta > 0$ and often $\theta > 1$) to correctly answer the majority of items pertaining to care of these maternal complications.

Overall, any provider would require an extremely high underlying competence level ($\theta > 2.8$ on a standard ±4 scale) to score 100%. We use Figure 1 to illustrate the levels of difficulty of items along with the levels of provider competence θ needed to succeed on vignette 1 items. On the whole, delivery care professionals in our sample would need θ in the neighbourhood of 1.4 (on ±4 scale) to correctly answer the full set of labour monitoring items. Those with θ in the vicinity of zero would fail the fifteenth item (routine examination of the placenta and membranes during the third stage labour), the second item (monitoring on cervical dilation during labour), and the tenth item (checking proteinuria of labouring women).

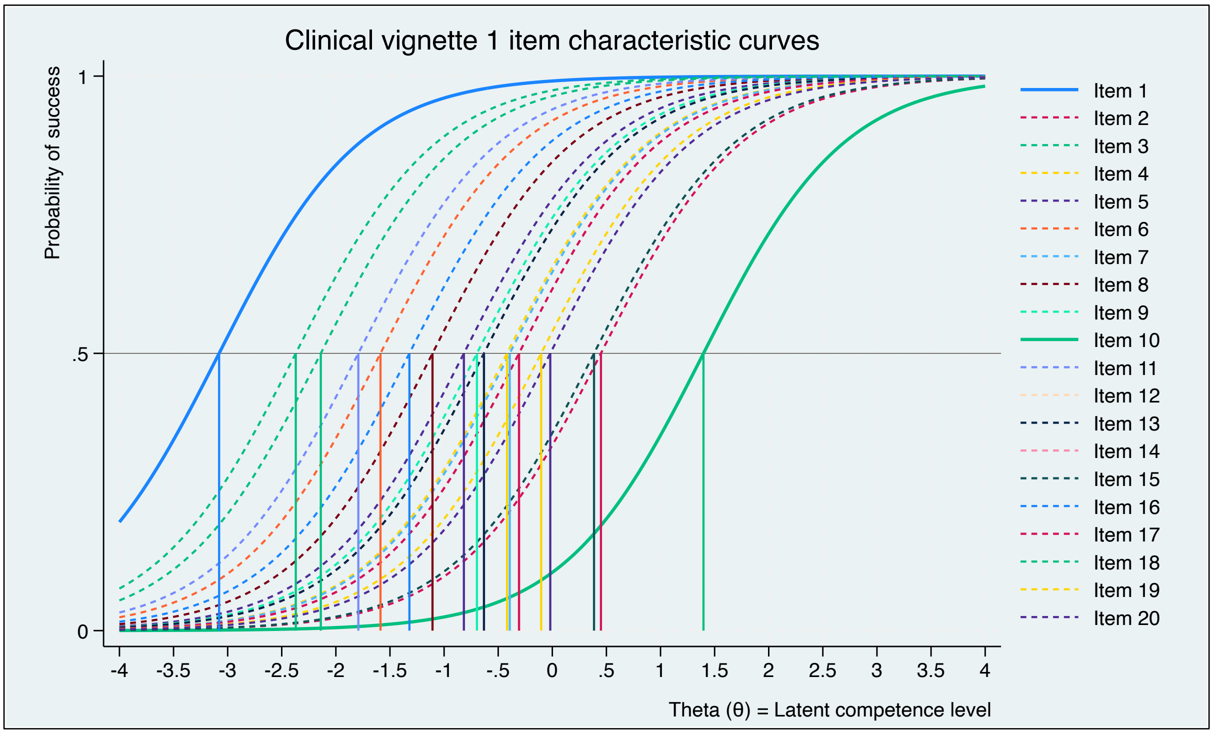


Figure 1 is the item characteristic curve of vignette 1 items (n=20). The horizontal axis represents the latent competence (θ) and the vertical axis is the probability of success on an item which is a function of both θ and the item difficulty b. The vertical lines represent the level of θ corresponding to 0.5 probability of success. Any level of θ below this threshold implies lower likelihood of success: if $\theta > b$; Pr(success = 1|0) —> 1 and if $\theta < b$; Pr(success = 1|0) —> 0. For example, professionals with θ in the vicinity of zero would succeed on 17/20 items situated on the left side of the vertical line $\theta = 0$ and fail item 15 (routine examination of the placenta and membranes during the third stage labour), item 2 (monitoring on cervical dilation during labour), and item 10 (checking proteinuria of labouring women) whose curves are on the right side of the vertical line $\theta = 0$.

Predicted EmONC competence test success rates

The expected success rates on the whole EmONC competence test if it were administered an infinite number of times are illustrated using Figure 2 and vignette-specific scores are summarised in Table 2. Keeping in mind that θ varies from —4 to +4, the combined analysis (Figure 2a) demonstrates that the above-average professionals (those with θ slightly above 0) are predicted to score 55.7 on the whole 115-item test, or an integer score of 55/115 or more items. Using the standard critical threshold for $\alpha = 0.05$, the graph demonstrates that we can expect 95% of randomly selected professionals in EmONC facilities of all qualifications combined to score between 14 and 98/115 items. Medical doctors (Figure 2b) with θ of 0 would pass 71/115 or more items (95% range 30 to 96 items) while average midwives (Figure 2c) and nurses (Figure 2d) are predicted to score 65 and 52/115 or more items, respectively. Regarding vignette-specific performance, Table 2 shows that the average professionals would score 13/20 items on labour monitoring, 4/13 items on fever and severe malaria management, 17/41 items on the management of postpartum haemorrhage, 8/24 items on the management of severe preeclampsia and eclampsia, and 10/17 items on the management of neonatal asphyxia and the provision of PMTCT services. Similar to the overall score, expected scores in individual vignettes consistently increase with higher qualification except for neonatal resuscitation and PMTCT in which all types of professionals seem to perform comparably. The predicted θ values in additional file 5 support the between profession differences with doctors having mean θ = 1.13 (SD = 0.88), midwives mean θ = 0.40 (SD = 0.84), and nurses mean θ = —0.12 (SD = 0.95).

Importantly, results in Table 2 reveal substantial gaps in EmONC competence within the same provider category corroborated by results of summated scores. For instance, we would expect some doctors to correctly perform only 2 of the 24 recommended clinical tasks to manage severe preeclampsia and eclampsia while others would handle up to 20 (83%) care items. Similarly, we would expect some nurses to fail the whole set of items relating to clinical management of severe preeclampsia and eclampsia while others would correctly achieve 19/24 of these care items.

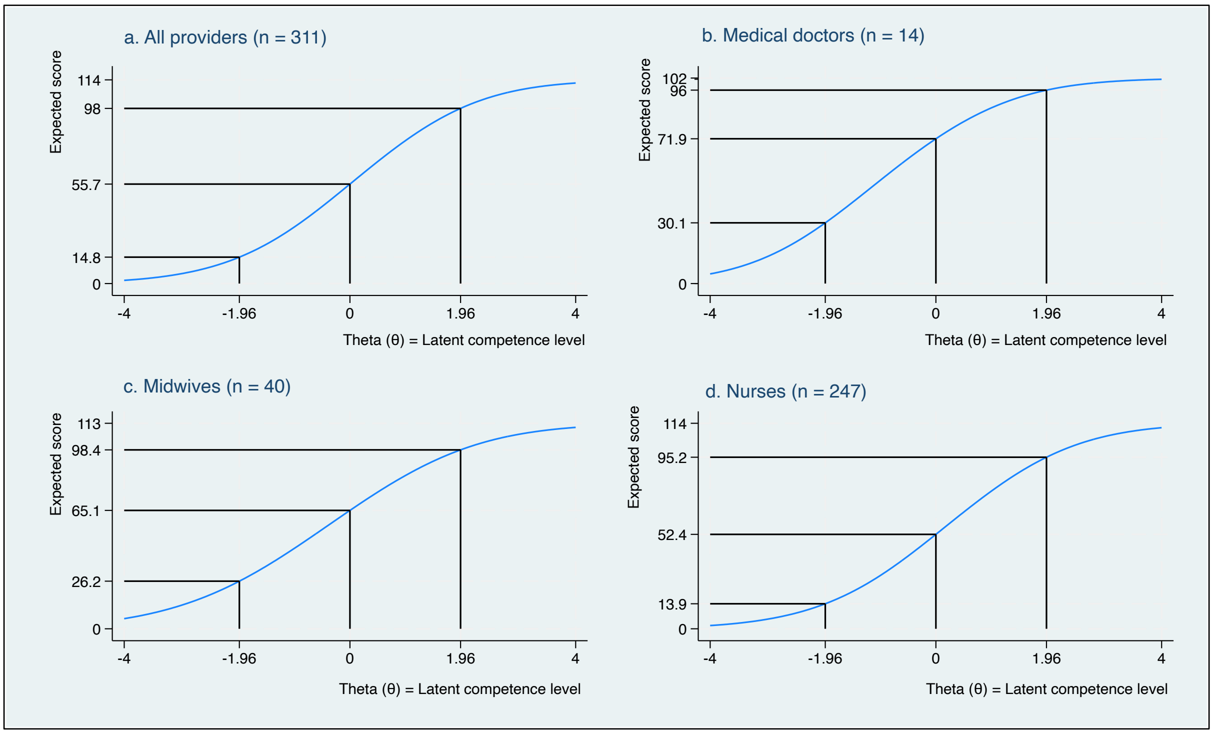


Figure 2 illustrates the Test Characteristic Curves (TCC) which predict the 'true score' as function of different levels of the latent competence (θ). The horizontal axis represents θ and the vertical axis is the predicted number of successes out of the total number of response items (n=115). We illustrate the predicted combined success rates (Figure 2a) and by provider type (Figures 2b-d) and use 95% critical values from the standard normal distribution (z = ±1.96) to predict the expected success rates of random samples of different types of deliver care professionals. We also predict the expected successes of averagely competent professionals (those with θ = 0).

	Percent mean score	**Predicted success rate**			
		Overall	Doctors	Midwives	Nurses
V 1. Woman in labour (n=20)	64.3 (34.4—94.1)	13 (4—19)	15 (3—17)	15 (8—19)	13 (4—19)
V 2. Fever and severe malaria (n=13)	36.5 (15.7—57.3)	4 (1—9)	7 (2—11)	4 (2—8)	4 (1—9)
V 3. Postpartum haemorrhage (n=41)	45.4 (23.5—67.3)	17 (2—33)	25 (7—35)	22 (5—36)	17 (2—33)
V 4. Severe preeclampsia or eclampsia (n=24)	35.6 (11.5—59.7)	8 (1—20)	14 (2—20)	12 (3—20)	7 (0—19)
V 5. Neonatal asphyxia and PMTCT (n=17)	58.6 (38.3—78.9)	10 (3—15)	11 (7—13)	11 (4—15)	10 (3—15)

Table 2. In the first column of the results section, we summarise the overall mean scores and the corresponding 95% CIs in parentheses obtained from raw scores. Subsequent columns summarise the predicted success rates for each vignette (overall and by provider type) for the averagely competent professionals (i.e. those with θ = 0) and the corresponding 95% CI success rates in parentheses assuming a countless repeat of the test administration.

Determinants of EmONC competence

Multivariable linear regression indicates that EmONC competence increases with higher qualification and with previous EmONC training (additional file 6). In a model that excluded doctors, the effect of facility type was not apparent suggesting midwives and nurses performed similarly in both CEmONC and BEmONC settings. Provider's sex and age, and the type of employment and length of maternity experience have no statistically significant effect on EmONC competence. Interestingly, EmONC training does not seem to significantly impact EmONC competence of medical doctors and midwives but does appear to have an effect in nurses (Table 3).

	Coefficient (95% CI)	S.E. (Bootstrap)	P-value
Midwives (base = nurses)	**0.52 (0.24 – 0.81)**	0.06	**0.000**
Doctors (base = nurses)	**1.25 (0.77 – 1.73)**	0.24	**0.000**
EmONC training (base = no training)	**0.33 (0.12 – 0.55)**	0.11	**0.002**
Doctors who received EmONC training†	0.27 (—0.65 – 1.18)	0.47	0.565
Midwives who received EmONC training†	—0.03 (—0.54 – 0.48)	0.26	0.905
Nurses who received EmONC training†	**0.37 (0.14 – 0.60)**	0.12	**0.002**

Table 3 summarises findings of the linear regression model examining the effects on qualification and EmONC training on true EmONC competence (predicted using the empirical Bayes means). We examined the effects of qualification and EmONC training separately and then combined by provider type. We used 10,000 bootstraps to address heteroskedasticity of residuals and highlighted significant coefficients in bold. Coefficients are θ and vary from -4 to +4. †Bases are professionals who did not receive EmONC training before.

EmONC competence test validity

Results of the test information function are demonstrated on additional file 7 Figure 1 which plots the information and standard error curves to diagnose how well the designed clinical vignettes are suited to estimating EmONC competence over the whole range of competence levels of delivery care professionals. Mathematically and graphically, the maximum information conveyed (i.e., the maximum precision) corresponds to the lowest standard errors. Inspection of this additional file 7 Figure 1 shows that the test gives the maximum information for θ in the neighbourhood of zero and remains high for the competence range $\theta \pm 0.5$. Within this range, EmONC competence examined using these clinical vignettes is estimated with greatest precision. Otherwise, the test validity decreases as the competence level departs from this range and the amount of information approaches zero at the extremes of the competence scale. Therefore, these clinical vignettes are suited for test consumers who have a competence level (θ) such that $-0.5 < \theta < 0.5$. In our sample and generally for Burundian delivery care professionals in EmONC facilities, 137/311 (44%) test takers of whom 1/14 (7%) medical doctor, 22/44 (55%) midwives, and 114/257 (44%) nurses would fall in that competence range and this implies that the instrument is believed to obtain precise EmONC competence estimates among midwives and nurses.

Discussion

Considering the growing impetus to capture competence of healthcare workers [6, 37] which is indispensable for achieving quality-adjusted service coverage and hence quality care [38], we show how a psychometric approach (i.e., the IRT) to examine competence of health workers can be used to deepen analysis compared to traditional and classical test theories (e.g., summated scales and component and factor analyses) that are the most popular forms of competence analysis in healthcare [15-17, 39, 40]. Classical approaches assume that test takers would achieve exactly the same scores if the test were to be administered repeatedly, fail to discern and discriminate between less and more difficult test components, and do not predict true competence which is essentially an unobservable trait [41, 42]. We used the EmONC competence test conducted on 311 delivery care providers across all Burundian EmONC maternities to demonstrate the applicability of the IRT model, a probabilistic model proposed by Rasch in 1960 [43], and how it overcomes the above limitations of classical test theories while providing a more comprehensive data appraisal [18]. While our main objective consisted in applying IRT to medical data rather than documenting its added value over existing tools, many authors have comprehensively explored and empirically proven this superiority [18, 23]. For instance, Kate Mitchell-Parker et al. (2018) demonstrated how various IRT models can enable precise measurement of individuals along the instrument by placing them and the test items on the same log-odds interval scale [18]. Moreover, a specific study that compared factor analysis against the IRT model has established that the latter model offers better fit in data applications involving individual rating particularly if there are many parameters and items [24].

In our study, we discovered that both raw statistics and the IRT model produce comparable learnings even though the IRT model offers several additional insights. This evidence is proven across results in Tables 1 to 5 of the additional file 2. They indicate that we would generally expect considerably smaller proportions of participants correctly respond to the most difficult test items (e.g., about 6% of participants would alert a medical specialist during the management of abundant postpartum haemorrhage or an incomplete/unsafe abortion which both have the highest b_j coefficients) while the vast majority of test takers would correctly answer the easiest items (e.g., 98% of participants would monitor fetal heart rate during childbirth and this item has a difficulty level b_j = —3.07 on ± 4 scale). Identically, the ANOVA test and the TCC and corresponding Bayes means estimates (additional file 7) both confirmed the effect of higher qualification on EmONC competence. Indeed, according to ANOVA findings (additional file 3), nurses achieved 44% on the EmONC competence test, midwives 56%, and doctors 70% and the TCC (Figure 2) emphasised these findings by further proving that random samples of doctors, midwives, and nurses would score 62% (71/115 successes), 57% (65/115 successes), and 45% (52/115 successes) respectively. Although these findings are similar, we argue that IRT predicted scores are most accurate as they draw on simulations of an infinite repeat of the EmONC test administration [22].

Furthermore, unlike ANOVA estimates which are percent scores based on correct answers, the Rasch model offered an extra layer of understanding of the latent EmONC competence which is fundamentally different from obtained scores and is estimated by updating prior information (i.e., the underlying EmONC competence) with the likelihood function (i.e., the actual test scores). Expectedly, we found substantial evidence in additional file 7 that medical doctors have higher EmONC competence than midwives (medical doctors' mean θ = 1.13, SD = 0.88 on θ ±4 scale versus mean θ = 0.40, SD = 0.84 among midwives) while nurses demonstrated the lowest competence level (mean θ = —0.12, SD = 0.95). The most policy relevant finding concerns the routine EmONC training which is heavily supported by the Burundian Ministry of Health (MoH) and key MNH partners principally UNFPA, WHO, UNICEF, the Japan International Cooperation Agency (JICA), Pathfinder, MNH professional associations, and various local research and service delivery institutions [44]. Its impact is notable among nurses as those who received it had on average about 0.37 (95% CI: 0.14 – 0.60) higher EmONC competence on θ = ±4 scale. Conversely, EmONC training does not have a statistically significant impact on the overall competence of medical doctors and midwives probably because the latter two professional categories receive substantial training in maternal and newborn care complemented with significant hospital internships in maternity and neonatology as part of their education.

Regarding the instrument validity, we proved that our EmONC clinical vignettes, although partly drawing on the standard EmONC training targeted at all delivery care professionals irrespective of their education levels, are more suitable to examine EmONC competence of nurses and midwives specifically. Factually, even though our sample included fewer medical doctors (14/311), we have evidence that this instrument would fail to detect precise EmONC competence of approximately 93% of random samples of Burundian medical doctors specifically those working in EmONC maternities. Therefore, a combination of the competence test results and the instrument validity suggests that medical doctors should receive a specifically designed EmONC competence test. Practically, nurses and probably midwives should receive more intense and prolonged training which has significant implications for the EmONC training strategy and budget.

Although we have extensively covered the most analytical approaches of the IRT model, there are some limitations to acknowledge. First, our focus on the one-parameter IRT model (i.e., the Rasch model which is limited to the difficulty parameter) didn't allow us to explore the discrimination and guessing parameters which is possible with a three-parameter IRT model. Second, our analysis relied on small sample sizes of medical doctors (14/311) and midwives (40/311) thus offering less information about these professionals, but this is a general depiction of Burundian healthcare workforce distribution and was improved by predicting an infinite number of test repeats that underpins the IRT model. Finally, the clinical vignettes did not test competence linked to obstetric surgery (e.g., hysterectomy and caesarean section) that might better discriminate between doctors. In the same perspective, we would therefore recommend conducting a specific study to examine doctors' competence linked to obstetric surgery.

Conclusion

This paper used a statistical method less commonly employed in health that rests upon the Bayesian concept to extend the common frequentist analyses of clinical competence of health workers. We demonstrated that IRT models reinforce findings of classical statistical theories namely summated scores while offering extra layers of understanding and enabling us to validate new or update existing competence test instruments tailored to particular categories of health workers. Particularly, we went beyond routine analytical approaches and illustrated the importance of capturing difficulty levels of various test components; proving the feasible computation of these analyses using commonly used and widely available statistical packages (e.g., STATA 18). Furthermore, we highlighted the major inherent assumptions underpinning competence assessment methods and showed how IRT concepts fully accommodate these assumptions. For instance, by predicting the true scores assuming an infinite repeat of test administration and by updating the latent competence (i.e., prior ability) using actual test scores (i.e., the likelihood function) to predict the true competence (i.e., posterior ability). We consistently found that Burundian delivery care professionals in EmONC facilities generally have poor EmONC competence although medical doctors proved to be better than midwives and nurses. We also found that Burundian standard EmONC training does not statistically augment EmONC competence of doctors and midwives and this has meaningful policy implications to redesign different EmONC training modules and competence assessment tools for medical doctors separately from midwives and nurses.

List of abbreviations

1PL	: One parameter logistic model
BEmONC	: Basic EmONC
CEmONC	: Comprehensive EmONC
EmONC	: Emergency obstetric and newborn care
HRH	: Human resources for health
ICC	: Item characteristic curve
IRT	: Item response theory
JICA	: Japan international cooperation agency
LMICs	: Low- and middle-income countries
MgSO4	: Magnesium sulphate
MNH	: Maternal and newborn health
MoH	: Ministry of Health
PMTCT	: Prevention of mother-to-child transmission of HIV

SDGs	: Sustainable Development Goals
TCC	: Total characteristic curve
TIF	: Test information function
UNFPA	: United Nations Population Fund
UNICEF	: United Nations Children's Fund
WHO	: World Health Organisation

Declarations

Acknowledgements

We acknowledge the members of the stakeholder committee in Burundi (additional file 1) and individual MNH experts namely Nadine Misago, Elvis Omondi Wambiya, Zacharie Ndizeye, Sandra Nkurunziza, Jean Paul Ndayizeye, Salvator Harerimana, Moise Muzigaba, Jalemba Aluvaala, Ambrose Agweyu, Kojo Nimako, and Abdu Adamu who shared expert opinion pertaining to the umbrella study that provided datasets used to perform these particular analyses.

Availability of data and files

This study used datasets from main EmONC facility survey and routine monitoring data which are not publicly available as primarily owned by the Burundian Ministry of Health (MoH) and are bound by a strong data sharing policy which does not permit authors to share them. However, these data can be obtained by sending a reasonable request to the reproductive, maternal, newborn, child, and adolescent's health programme of the Burundian MoH. The corresponding author can share STATA command files upon request.

Competing interests

None declared.

Authors' contributions

DH conceptualized and wrote the study protocol, engaged policy makers and stakeholders in Burundi, sought ethics approvals and WHO funding for fieldwork, curated and analysed data, and wrote the first and final drafts of the manuscript. AL, CN, and ME contributed substantially to the study conceptualization, guided data analysis, and reviewed the drafts manuscript. JBN, AN, AB, JN, ESDN, and SB reviewed the study protocol and helped to coordinate stakeholder discussions. PN† contributed to the study conceptualisation but died during the course of the study conduct. All authors read and approved the final manuscript.

Ethics approval and consent to participate

This study was approved by the Human Research Ethics Committee of the Faculty of Medicine of the University of Burundi (Ethics certificate FM/CE/01/M/2022) and the University of Oxford's Tropical Research Ethics Committee (OxTREC approval reference: 516-22). Participation was voluntary and all participants signed a written informed consent form.

Funding

Data collection activities were funded by WHO country office in Burundi. Mike English contributed expertise thanks to his Wellcome Trust Senior Research Fellowship (#207522). None of these funders had a role in data analysis and reporting or in deciding to publish this manuscript. The authors had the final responsibility to decide to submit the manuscript for publication.

Consent for publication

Not applicable.

References

1. Burundi, *Ministère à la Présidence chargé de la Bonne Gouvernance et du Plan (MPBGP), Ministère de la Santé Publique et de la Lutte contre le Sida (MSPLS), Institut de Statistiques et d'Études Économiques du Burundi (ISTEEBU), et ICF. Troisième Enquête Démographique et de Santé. Bujumbura, Burundi : ISTEEBU, MSPLS, et ICF*. 2017.
2. Say, L., et al., *Global causes of maternal death: a WHO systematic analysis.* The Lancet global health, 2014. **2**(6): p. e323-e333.
3. WHO, *World Health Organization: Strategies towards ending preventable maternal mortality (EPMM).* 2015.
4. Willcox, M.L., et al., *Human resources for primary health care in sub-Saharan Africa: progress or stagnation?* Human resources for health, 2015. **13**: p. 1-11.
5. Lucas, A.O., *Human resources for health in Africa*. 2005, British Medical Journal Publishing Group. p. 1037-1038.
6. Kak, N., B. Burkhalter, and M.-A. Cooper, *Measuring the competence of healthcare providers.* Operations research issue paper, 2001. **2**(1): p. 1-28.
7. Khan, A.N.S., et al., *Competence of healthcare professionals in diagnosing and managing obstetric complications and conducting neonatal care: a clinical vignette-based assessment in district and subdistrict hospitals in northern Bangladesh.* BMJ open, 2019. **9**(8): p. e028670.
8. Puri, R., et al., *Knowledge, attitudes, and practices in safe motherhood care among obstetric providers in Bugesera, Rwanda.* International Journal of Gynecology & Obstetrics, 2012. **116**(2): p. 124-127.
9. Kim, Y.M., et al., *Assessing the capacity for newborn resuscitation and factors associated with providers' knowledge and skills: a cross-sectional study in Afghanistan.* BMC pediatrics, 2013. **13**(1): p. 1-12.
10. Mirkuzie, A.H., et al., *Current evidence on basic emergency obstetric and newborn care services in Addis Ababa, Ethiopia; a cross sectional study.* BMC pregnancy and childbirth, 2014. **14**(1): p. 1-8.
11. Das, J. and J.S. Hammer, *Which Doctor?: Combining Vignettes and Item-response to measure Doctor Quality*. Vol. 3301. 2004: World Bank Publications.
12. Spencer, L.M., D.C. McClelland, and S.M. Spencer, *Competency assessment methods: History and state of the art.* 1994.
13. Blouse, A., *Guidelines for assessment of skilled providers after training in maternal and newborn health care.* United States of America: JHPIEGO, 2004.
14. Landry, E., et al., *Assessing the quality of record keeping for cesarean deliveries: results from a multicenter retrospective record review in five low-income countries.* BMC pregnancy and childbirth, 2014. **14**: p. 1-11.
15. Huchon, C., et al., *Obstetric competence among referral healthcare providers in Mali.* International Journal of Gynecology & Obstetrics, 2014. **126**(1): p. 56-59.
16. Gupta, M., et al., *Competency assessment of the medical interns and nurses and documenting prevailing practices to provide family planning services in teaching hospitals in three states of India.* PLoS One, 2019. **14**(11): p. e0211168.

17. Ameh, C.A., et al., *Knowledge and skills of healthcare providers in sub-Saharan Africa and Asia before and after competency-based training in emergency obstetric and early newborn care.* PloS one, 2016. **11**(12): p. e0167270.

18. Mitchell-Parker, K., et al., *Rasch analysis of the frost multidimensional perfectionism scale.* Australian Journal of Psychology, 2018. **70**(3): p. 258-268.

19. Baker, F., *The basics of item response theory. Chapter 4: The Test Characteristic Curve*. 2001, University of Wisconsin, ISBN 1-886047-03-0: ERIC Clearinghouse on Assessment and Evaluation.

20. Lord, F.M., *Practical applications of item characteristic curve theory.* Journal of Educational Measurement, 1977: p. 117-138.

21. Molenaar, D., *Review of Handbook of Item Response Theory, Volume II: Statistical Tools*. 2020, SAGE Publications Sage CA: Los Angeles, CA.

22. Das, J. and J. Hammer, *Which doctor? Combining vignettes and item response to measure clinical competence.* Journal of Development Economics, 2005. **78**(2): p. 348-383.

23. Tennant, A. and P.G. Conaghan, *The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper?* Arthritis Care & Research, 2007. **57**(8): p. 1358-1362.

24. Hayat, B., M.D.K. Putra, and B. Suryadi, *Comparing item parameter estimates and fit statistics of the Rasch model from three different traditions.* Jurnal Penelitian dan Evaluasi Pendidikan, 2020. **24**(1): p. 39-50.

25. Cai, L., et al., *Item response theory.* Annual Review of Statistics and Its Application, 2016. **3**: p. 297-321.

26. Reise, S.P. and N.G. Waller, *Item response theory and clinical measurement.* Annual review of clinical psychology, 2009. **5**: p. 27-48.

27. Thorpe, G.L. and A. Favia, *Data Analysis Using Item Response Theory Methodology: An Introduction to Selected Programs and Applications.* 2012.

28. Habonimana, D., et al., *Eight years into the horizon of aspirational maternal and newborn health pledges: a nationwide cross-sectional exploration of the Burundian EmONC network capacity and budget deficits.* BMJ Open, 2024. **14**(5): p. e083546.

29. Habonimana, D., et al., *Addressing the need for an appropriate skilled delivery care workforce in Burundi to support Maternal and Newborn Health Service Delivery Redesign (MNH-Redesign): a sequential study protocol.* Wellcome Open Research, 2022. **7**.

30. Traoré, M., et al., *Obstetric competence among primary healthcare workers in Mali.* International Journal of Gynecology & Obstetrics, 2014. **126**(1): p. 50-55.

31. WHO, *Managing complications in pregnancy and childbirth: a guide for midwives and doctors – 2nd ed. Geneva: World Health Organization. Licence: CC BY-NC-SA 3.0 IGO.* 2017.

32. Baker, F.B. and S.-H. Kim, *The basics of item response theory using R*. DOI 10.1007/978-3-319-54205-8. Vol. 969. 2017: Springer.

33. Lawley, D.N., *XXIII.—On problems connected with item selection and test construction.* Proceedings of the Royal Society of Edinburgh Section A: Mathematics, 1943. **61**(3): p. 273-287.

34. Hambleton, R.K. and H. Swaminathan, *Item response theory: Principles and applications*. 2013: Springer Science & Business Media.

35. Baker, F.B., et al., *The test characteristic curve.* The Basics of Item Response Theory Using R, 2017: p. 55-67.

36. Cribari-Neto, F. and S.G. Zarkos, *Bootstrap methods for heteroskedastic regression models: evidence on estimation and testing.* Econometric Reviews, 1999. **18**(2): p. 211-228.

37. Lin, C.-J., C.-K. Lee, and M.-C. Huang, *Cultural competence of healthcare providers: A systematic review of assessment instruments.* Journal of Nursing Research, 2017. **25**(3): p. 174-186.

38. Amouzou, A., et al., *Advances in the measurement of coverage for RMNCH and nutrition: from contact to effective coverage.* BMJ global health, 2019. **4**(Suppl 4): p. e001297.

39. Smith, V., K. Muldoon, and L. Biesty, *The Objective Structured Clinical Examination (OSCE) as a strategy for assessing clinical competence in midwifery education in Ireland: A critical review.* Nurse education in practice, 2012. **12**(5): p. 242-247.

40. Vincent, S.C., et al., *Objective structured clinical examination vs traditional clinical examination to evaluate students' clinical competence: A systematic review of nursing faculty and students' perceptions and experiences.* Nurse Education Today, 2022. **108**: p. 105170.

41. Rahayu, W., et al., *A Rasch and factor analysis of an Indonesian version of the Student Perception of Opportunity Competence Development (SPOCD) questionnaire.* Cogent Education, 2020. **7**(1): p. 1721633.

42. Heuling, L.S., S. Wild, and A. Vest, *Digital competences of prospective engineers and science teachers: A latent profile and correspondence analysis.* 2021.

43. Rasch, G., *Probabilistic models for some intelligence and attainment tests*. 1993: ERIC.

44. Burundi, *Ministère de la santé publique et de lutte contre le SIDA. Évaluation des besoins en matière de soins obstétricaux et néonatals d'urgence au Burundi « EB SONU ».* Institut de Statistiques et d'Etudes Economiques du Burundi (ISTEEBU). 2011: p. Page 29.

Additional file 1 : List of study stakeholders

Names	Position and institution
1. Dr Jean Baptiste Nzorironkankuze	Permanent Secretary, Ministry of Health (MoH)
2. Dr Oscar Ntihabose	Director General Health Services, MoH
3. Dr Ananie Ndacayisaba	Director of the RMNCAH Programme, MoH
4. Dr Zacharie Kubwimana	Director of Programme, MoH
5. Dr Theophile Bigayi	Deputy Director RMNCAH Programme, MoH
6. Prof Sylvestre Bazikamwe*	Professor of Gynecology and Obstetrics , University of Burundi
7. Dr Jeanine Ndayisenga**	Paediatrician Research Fellow, National Institute of Public Health
8. Dr Yolande Magonyagi	National Programme Officer SRH, United Nations Population Fund (UNFPA)
9. Dr Eugenie Siga Diane Niane	Programme Officer RMNCAH, World Health Organization (WHO)
10. Dr Brigitte Ndelema	National Programme Officer RMNCAH, WHO
11. Dr Aristide Bishinga	EmONC Specialist, Japan International Cooperation Agency (JICA)
12. Dr Dorothee Ntakirutimana	Health Specialist, United Nations Children's Fund (UNICEF)
13. Dr Josiane Nijimbere	Scientist RMNCAH Programme, MoH
14. Dr Jean Claude Mugisha	Scientist RMNCAH Programme, MoH
15. Dr Innocent Nkurunziza	Secretary, MoH
16. Dr Anaclet Nahayo	Director of HIS, MoH
17. Olivier Gahungere, Mr	Statistician, RMNCAH Programme, MoH
18. Thierry Nzeyimana, Mr	Monitoring and Evaluation, RMNCAH Programme, MoH
19. Daniel Habonimana, Mr	Monitoring and Evaluation, RMNCAH Programme, MoH
20. Souverienne Bucumi, Mrs	Monitoring and Evaluation, RMNCAH Programme, MoH
21. Rose Simone Ndayiziga, Mrs	Scientist, National HIV/AIDS programme, MoH
22. Nadine Muhimbare, Ms.	Cabinet Secretary, MoH
23. Therence Nduwarugira, Mr	Scientist, UNFPA

*He also represents the Burundian Association of Gynecology-Obstetricians (AGOB)
**She also represents the Burundian Association of Neonatology (ABUNE)

Additional file 2 Table 1. Difficulty levels of clinical vignette 1 items

Question	Item number and correct practice	Professionals with correct answers (n, %)	Item difficulty level (b_j)
What are the key elements to monitor regularly during childbirth?	1. Fetal heart rate	304 (97.75)	−3.07
	2. Cervical dilatation	293 (94.21)	−2.37
	3. Mother's blood pressure	266 (85.53)	−1.59
	4. Mother's temperature	239 (76.85)	−1.11
	5. Uterine contractions	219 (70.42)	−0.82
	6. Mother's pulse	210 (67.52)	−0.71
	7. Head descent	188 (60.45)	−0.42
	8. Mother's breathing rate	186 (59.81)	−0.39
	9. Fetal head position	117 (37.62)	0.44
	10. Urine protein	179 (57.56)	1.39
What are the components of the active management of the third stage of labour?	11. Oxytocin injection	287 (92.28)	−1.79
	12. Controlled cord traction	162 (52.09)	−0.63
	13. Uterine massage after delivery of the placenta	155 (49.84)	−0.63
	14. Routine examination of the vagina and perineum	55 (17.68)	−0.39
	15. Examination of the placenta and membranes	304 (97.75)	0.31
What are the safety measures to undertake to prevent septic childbirth?	16. Wear sterile gloves on both hands	179 (57.56)	−2.14
	17. Put on a clean plastic apron, boots and glasses	275 (88.42)	−1.32
	18. Wash hands with antiseptic soap, dry with a sterile cloth	205 (65.92)	−0.31
	19. Clean the perineum with a compress soaked in antiseptic solution, wiping from front to back	122 (39.23)	−0.10
	20. Place three sterile drapes, one under the woman's buttocks, one over her abdomen, and the other one reserved to receive the newborn	252 (81.03)	−0.02

Additional file 2 Table 1. Labour monitoring and infection prevention during childbirth

Additional file 2 Table 2. Difficulty levels of clinical vignette 2 items

Question	Response items	Professionals with correct answers (n, %)	Item difficulty level (b_j)
What are the actions to do when a pregnant woman has severe malaria symptoms?	21. Quinine and dextrose IV	252 (81.03)	−1.76
	22. Proceed immediately with admission	215 (69.13)	−0.99
	23. Start the vital signs chart	179 (57.56)	−0.38
	24. Check haemoglobin	133 (42.77)	0.35
	25. Blood slide for MPS	112 (36.01)	0.70
What are the danger signs of severe malaria during pregnancy?	26. Jaundice (Yellowish eyes)	134 (43.09)	0.33
	27. Dizziness	119 (38.26)	0.58
	28. Fever (above 38°C)	93 (29.90)	1.04
	29. Pallor	74 (23.79)	1.42
	30. Condition of the foetus	68 (21.86)	1.55
	31. Joint pain	64 (20.58)	1.64
	32. Confusion/Coma	55 (17.68)	1.87
	33. Dehydration	55 (17.68)	1.87

Additional file 2 Table 2. Management of fever and malaria during pregnancy

Additional file 2 Table 3. Difficulty levels of clinical vignette 3 items

Question	Response items	Professionals with correct answers (n, %)	Item difficulty level (b_j)
What are the danger signs to monitor in case of abundant maternal bleeding?	34. Signs of shock (dizziness, low BP)	220 (70.74)	−0.86
	35. Genital tract injuries	218 (70.10)	−0.84
	36. Retained products/placenta	213 (68.49)	−0.77
	37. Signs of anaemia	193 (62.06)	−0.52
	38. Uncontracted uterus	165 (53.05)	−0.18
	39. Amount of external bleeding	141 (45.34)	0.11
	40. Check if bladder is full	48 (15.43)	1.59
What are the important and systematic actions to do in case of abundant maternal bleeding?	41. Give ergometrine IM or IV (Oxytocin)	225 (72.35)	−0.93
	42. Examine the woman for lacerations	217 (69.77)	−0.82
	43. Manual removal of retained products	202 (64.95)	−0.63
	44. Start IV fluids	199 (63.99)	−0.59
	45. Massage the fundus	130 (41.80)	0.25
	46. Take blood for haemoglobin and crossmatching	129 (41.48)	0.26
	47. Empty the woman's bladder	66 (21.22)	1.20
	48. Raise foot of bed	66 (21.22)	1.20
	49. Inform a medical specialist	20 (6.43)	2.53
What are the key and systematic actions to do in case of a retained placenta?	50. Give IV fluids	239 (76.85)	−1.14
	51. Apply manual removal of the placenta	200 (64.31)	−0.60
	52. Repeat oxytocics	114 (36.66)	0.46
	53. Test blood for group and cross match	100 (32.15)	0.65
	54. Check uterus is well contracted	93 (29.90)	0.75
	55. Monitor vital signs for shock & act	84 (27.01)	0.89
	56. Empty urinary bladder	60 (19.29)	1.32
	57. Apply controlled cord traction	49 (15.76)	1.56
	58. Inform a medical specialist	39 (12.54)	1.83
	59. Prepare for theatre	35 (11.25)	1.95
What are the signs to monitor when a mother develops malaise within 24 hours after delivery?	60. Septic shock (unrecordable blood pressure)	249 (80.06)	−1.31
	61. Tender abdomen	182 (58.52)	−0.38
	62. High pulse	179 (57.56)	−0.35
	63. Sensitive underinvoluted uterus	159 (51.13)	−0.11
	64. Foul smelling lochia	136 (43.73)	0.17
	65. High fever	98 (31.51)	0.68
What are the key actions to undertake following an incomplete or an unsafe abortion?	66. Do manual vacuum aspirations (MVA)	190 (61.09)	−0.48
	67. Start on antibiotics	187 (60.13)	−0.44
	68. Assess the vital signs	186 (59.81)	−0.43
	69. Assess the bleeding	155 (49.84)	−0.06
	70. Do vaginal examination	153 (49.20)	−0.03
	71. Do conventional evacuation (D&C)	105 (33.76)	0.58
	72. Start on IV fluids	85 (27.33)	0.87
	73. Provide counselling	64 (20.58)	1.24
	74. Inform a medical specialist	19 (6.11)	2.58

Additional file 2 Table 3. Management of postpartum haemorrhage

Additional file 2 Table 4. Difficulty levels of clinical vignette 4 items

Question	Response items	Professionals with correct answers (n, %)	Item difficulty level (b_j)
What to do when a mother presents severe headaches and abdominal pain?	75. Take vital signs (especially temperature and BP)	272 (87.46)	—1.59
	76. Check the term of pregnancy	163 (52.41)	—0.12
	77. Get details about the abdominal pain (epigastric)	106 (34.08)	0.52
	78. Confirm any visual problems (blurred vision)	76 (24.44)	0.93
What are the positive signs to check to confirm suspected severe preeclampsia?	79. High level of BP (diastolic > 100mmhg)	256 (82.32)	—1.29
	80. Severe headache	213 (68.49)	—0.69
	81. Proteinuria (3+)	192 (61.74)	—0.44
	82. Visual disturbance	132 (42.44)	0.22
	83. Epigastric pain	116 (37.30)	0.40
What do you monitor during severe preeclampsia treatment?	84. Monitor vital signs, reflexes and FHR hourly	209 (67.20)	—0.64
	85. Test proteinuria	90 (28.94)	0.73
	86. Maintain a strict fluids balance chart and monitor the amount of fluids intake to avoid fluid overload	87 (27.97)	0.76
	87. Monitor for possible pulmonary oedema	60 (19.29)	1.20
	88. Catheterize the bladder to monitor urine output	59 (18.97)	1.21
	89. Assess clotting to rule out coagulopathy	25 (8.04)	2.04
Severe preeclampsia is confirmed, what actions to be undertaken?	90. Continue treatment with Mgso4 for 24 hours after delivery or convulsion (whichever is late)	106 (34.08)	0.53
	91. Give 20% magnesium sulfate solution 4g IV, over 5 minutes	87 (27.97)	0.78
	92. Follow promptly with 50% magnesium sulfate solution 5g, IM in each buttocks with 2% lignocaine 1ml in the same syringe	66 (21.22)	1.09
	93. After every 4 hours, continue the maintenance dose of 50% magnesium sulfate solution 5g; with 2% lignocaine 1ml in the same syringe, in IM into alternate buttocks	50 (16.08)	1.39
Precautions to observe before administering the maintenance dose of MgSO4	94. Respiratory rate should be at least 16 cycles per minute	0 (00)	n.a.
	95. Urinary output is at least 30ml/hour over 4 hours	66 (21.22)	1.09
	96. Patellar reflexes are present	46 (14.79)	1.47
Actions for cardiac or pulmonary arrest following MgSO4 administration	97. Assisted ventilation	99 (31.83)	0.62
	98. Give Calcium gluconate 1g (or 10ml of 10% solution) IV slowly until respiration begins to antagonise the effect of MgSO4	136 (43.73)	0.17

Additional file 2 Table 4. Management of severe preeclampsia and eclampsia

Additional file 2 Table 5. Difficulty levels of clinical vignette 5 items

Question	Response items	Professionals with correct answers (n, %)	Item difficulty level (b_j)
What are the systematic elements of newborn care immediately after birth?	99. Cord care (sterile cut 4–6 cm/umbilicus)	291 (93.57)	−2.64
	100. After birth of head, wipe face, nose, mouth	241 (77.49)	−1.29
	101. Assess/examine newborn within one hour	213 (68.49)	−0.83
	102. Ensure baby is breathing	200 (64.31)	−0.64
	103. Eye prophylaxis	191 (61.41)	−0.51
	104. Breastfeeding initiated within one hour	161 (51.77)	−0.09
	105. Thermal protection	132 (42.44)	0.30
	106. Baby weighed	115 (36.98)	0.54
What are the essential and systematic steps of birth asphyxia treatment?	107. Clear the airways of blood and mucus	244 (78.46)	−1.35
	108. Assist breathing with ambu bag	239 (76.85)	−1.26
	109. Provide suction if there is meconium	210 (67.52)	−0.78
	110. Dry and wrap baby in a warm clean cloth	169 (54.34)	−0.20
	111. Do cardiac massage	160 (51.45)	−0.08
	112. Place under radiant heat warmer	122 (39.23)	0.44
What is the systematic PMTCT protocol?	113. Provide NVP to infant	278 (89.39)	−2.15
	114. Offer testing for those with unknown serostatus	144 (46.30)	0.14
	115. Provide NVP in early labour for HIV+ woman	68 (21.86)	1.32

Additional file 2 Table 5. Immediate care for the newborn

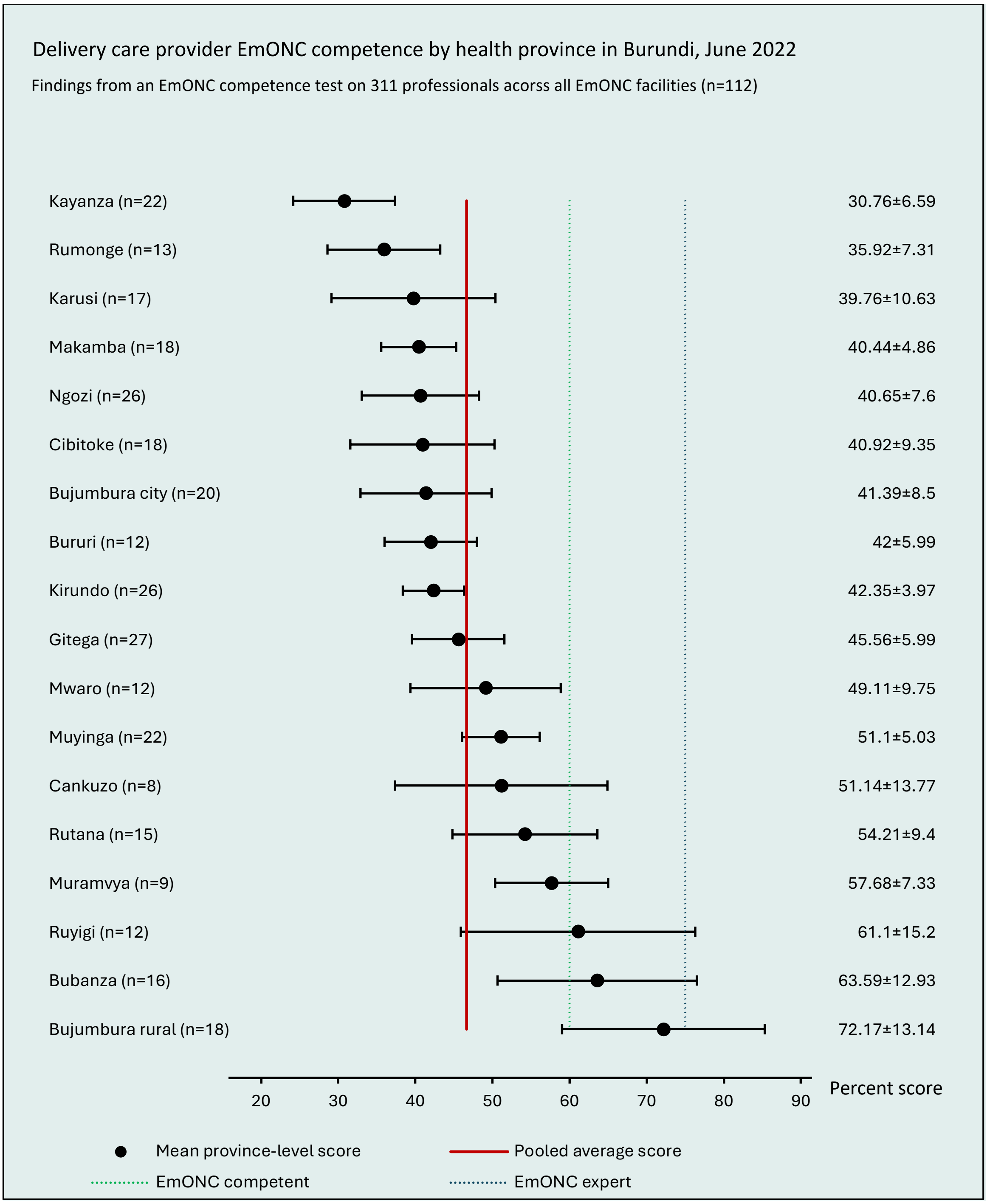
Additional file 3. ANOVA findings

EmONC clinical skills questions	Items description			Percent mean scores (SD)				Group difference	
	Number and sign	Inter-item correlation	Cronbach's alpha	Overall (n=311)	Nurses (n=257)	Midwives (n=40)	Doctors (n=14)	*F* statistic	p-value
Labour monitoring (n=2)	**15, *p****	**0.59**	**0.7384**	**65.1 (25.5)**	**61.6 (25.7)**	**81.0 (14.2)**	**84.3 (22.3)**	**15.5**	**0.0000**
What are the key elements to monitor during childbirth?	10, *p**	0.22	0.7384	66.8 (22.9)	63.9 (23.2)	78.5 (14.4)	85.7 (17.4)	12.9	0.0000
What are the indications and key components of the active management of third stage of labour?	5, *p**	0.45	0.8032	63.40(34.0)	59.2 (34.5)	83.5 (20.2)	82.9 (30.2)	11.9	0.0000
Infection prevention during childbirth (n=1)	**5, *p****	**0.13**	**0.4235**	**63.4 (34.1)**	**59.2 (34.5)**	**83.5 (20.2)**	**82.9 (30.2)**	**5.9**	**0.0032**
What are the safety measures to undertake for preventing septic childbirth?	5, *p**	0.13	0.4235	63.4 (34.1)	59.2 (34.5)	83.5 (20.2)	82.9 (30.2)	5.9	0.0032
Management of malaria during pregnancy (n=2)	**13, *p****	**0.13**	**0.2299**	**36.5 (20.8)**	**35.7 (20.3)**	**34.6 (18.5)**	**59.4 (23.6)**	**9.4**	**0.0001**
What are the immediate actions when a pregnant women present severe malaria symptoms?	5, *p**	0.35	0.7249	28.3 (30.3)	28.1 (29.4)	24.5 (31.5)	42.9 (39.1)	2.0	0.1435
What are the danger signs of severe malaria during pregnancy?	8, *p**	0.21	0.6759	44.7 (24.9)	30.4 (24.2)	44.7 (22.4)	75.9 (25.7)	12.4	0.0000
Management of postpartum haemorrhage (n=5)	**41, *p****	**0.69**	**0.9182**	**45.4 (21.9)**	**42.8 (21.1)**	**53.7 (19.9)**	**70.3 (20.0)**	**15.1**	**0.0000**
What are the danger signs to monitor in case of abundant maternal bleeding?	7, *p**	0.18	0.6108	55.0 (25.3)	52.2 (24.8)	63.9 (22.4)	81.6 (19.8)	12.8	0.0000
What are the important and systematic actions to do in case of abundant maternal bleeding?	9, *p**	0.20	0.6899	44.8 (23.4)	42.0 (22.5)	54.2 (22.9)	69.0 (21.9)	13.6	0.0000
What are the key and systematic actions to do in case of a retained placenta?	10, *p**	0.28	0.7947	32.6 (24.8)	30.1 (23.6)	41.3 (25.40	52.8 (31.0)	8.8	0.0002
What are the signs to monitor when a mother develops malaise within 48 hours after delivery?	6, *p**	0.20	0.6062	53.8 (27.6)	51.2 (27.2)	60.0 (25.8)	82.1 (23.1)	10.1	0.0001
What are the key actions to undertake following an incomplete or an unsafe abortion?	9, *p**	0.22	0.7132	40.9 (25.0)	38.2 (24.4)	49.2 (23.3)	65.9 (23.2)	11.4	0.0000
Management of preeclampsia & eclampsia (n=6)	**24, *p****	**0.53**	**0.8706**	**35.6 (24.1)**	**31.7 (22.8)**	**50.2 (19.9)**	**64.5 (23.7)**	**23.8**	**0.0000**
What do you suspect when a mother presents severe headaches and abdominal pain?	4, *p**	0.25	0.5715	49.6 (29.3)	47.2 (29.0)	53.8 (25.0)	82.1 (24.9)	10.6	0.0000
What are the positive signs to check to confirm suspected severe preeclampsia?	5, *p**	0.27	0.6484	58.5 (30.2)	53.6 (30.0)	79.5 (21.4)	87.1 (18.6)	21.9	0.0000
What are the essential monitoring elements in case of severe preeclampsia?	6, *p**	0.35	0.7615	28.4 (27.2)	25.4 (26.0)	37.1 (24.6)	59.5 (32.5)	13.9	0.0000

What are the important and systematic actions to do after confirmation of severe preeclampsia?	4, *p**	0.38	0.7105	16.0 (24.6)	12.5 (22.0)	29.0 (28.6)	44.0 (29.9)	19.4	0.0000
What are the key elements to assess before repeat administration of magnesium sulfate?	3, *p**	0.47	0.7307	23.4 (33.3)	19.2 (31.1)	40.8 (34.2)	50.0 (42.9)	12.9	0.0000
What actions to take in case of a cardiopulmonary arrest after the administration of MgSO4?	2, *p**	0.36	0.5269	37.8 (39.7)	32.7 (38.6)	61.3 (36.7)	64.3 (36.3)	13.2	0.0000
Immediate care for the newborn (n=3)	**17, *p****	**0.46**	**0.7193**	**58.6 (20.3)**	**57.4 (20.8)**	**61.9 (17.6)**	**72.4 (11.3)**	**4.3**	**0.0141**
What are the systematic elements of newborn care immediately after birth?	8, *p**	0.22	0.6944	62.1 (26.0)	60.7 (26.4)	65.6 (23.6)	76.8 (20.7)	3.01	0.0506
What are the essential and systematic steps of birth asphyxia treatment?	6, *p**	0.19	0.5769	61.3 (26.5)	58.8 (26.4)	70.0 (23.0)	83.3 (23.6)	8.58	0.0002
What is the systematic approach of PMTCT?	3, *p**	0.16	0.3705	52.5 (23.2)	52.7 (23.6)	50.0 (20.0)	57.1 (24.2)	0.52	0.5965
Overall score (n=19)	**115, *p****	**0.48**	**0.9452**	**46.6 (19.3)**	**44.0 (18.9)**	**55.7 (15.6)**	**69.9 (16.7)**	**18.7**	**0.0000**

Additional file 3 contains results from the clinical skills exam. In total, the exam comprised six skill subsets, 19 questions, and 115 response items. The items correlated positively internally within each question and externally between questions and skill subsets. Reliability tests proved internal consistency across question and skill subset items. We reported the overall percent mean score on each question, each skill subset, and the overall score on the whole exam. Results are presented for all respondents (column 4) and for each type of provider (columns 5-7). We also reported significance level of the mean score differences between professionals (column 9).

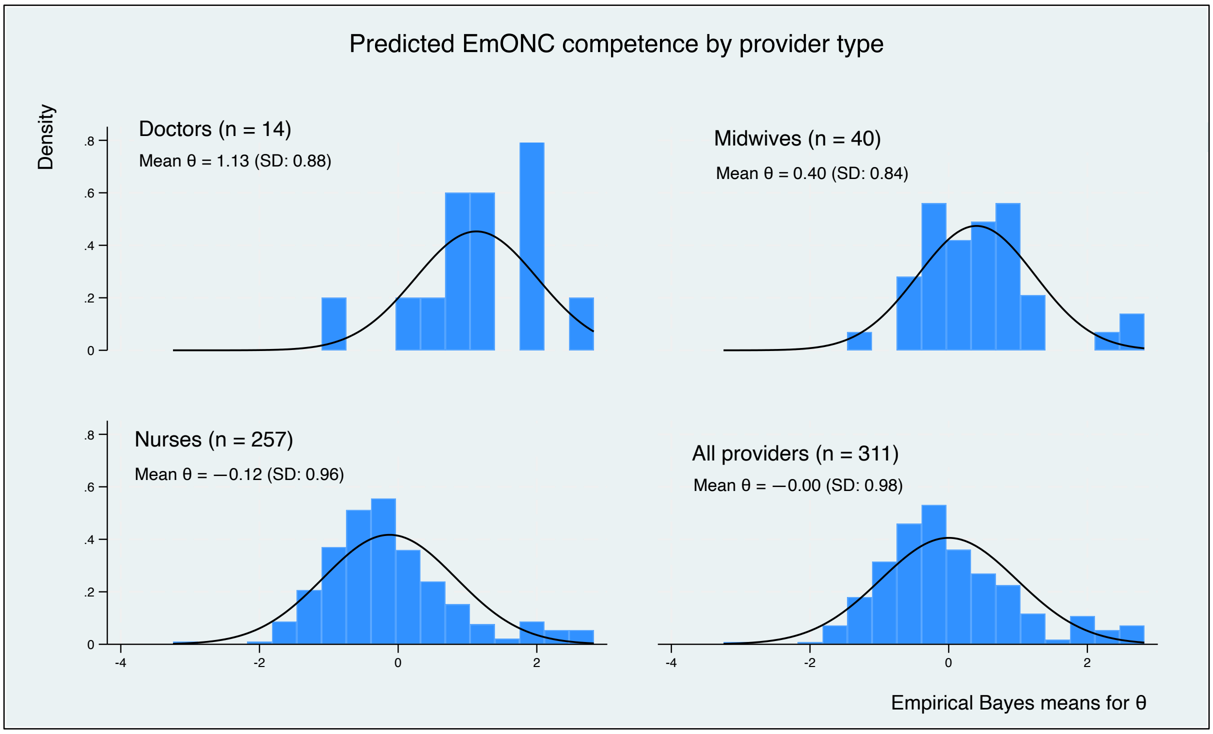
Additional file 4. EmONC competence test scores by province



Additional file 4 illustrates the average EmONC competence test scores at provincial level using total raw scores (out of 115 items corresponding to 115 maximum possible score). The Figure demonstrates provinces with poor EmONC performance in descending order. We highlight the national-level average score in red (pooled average score is around 48%) and provinces that achieved 60% (competent level) and 75% (expert level) on this EmONC competence test.

Additional file 5. Predicted competence level by provider category

To obtain the EmONC competence of each delivery care provider, we used the empirical Bayes means estimates which combine prior information about the latent competence θ with the likelihood function to obtain the conditional posterior θ distribution on ±4 scale. A comparison of true EmONC competence between delivery care professional groups is illustrated in additional file 5 Figure 1 and further substantiates the increase of EmONC competence with higher qualification. Expectedly, half of medical doctors have $\theta > 1.25$ (mean $\theta = 1.13$, SD = 0.88) while half of midwives have $\theta = 0.35$ (mean $\theta = 0.40$, SD = 0.84). Nurses have the lowest competence level with median $\theta =$ —0.25 (mean $\theta =$ —0.12, SD = 0.95). Interestingly, the overall mean competence appears to be significantly lower (mean $\theta =$ —0.00, SD = 0.98) compared to those of medical doctors and midwives separately and this is due to the predominant presence of nurses in our sample (257/311).



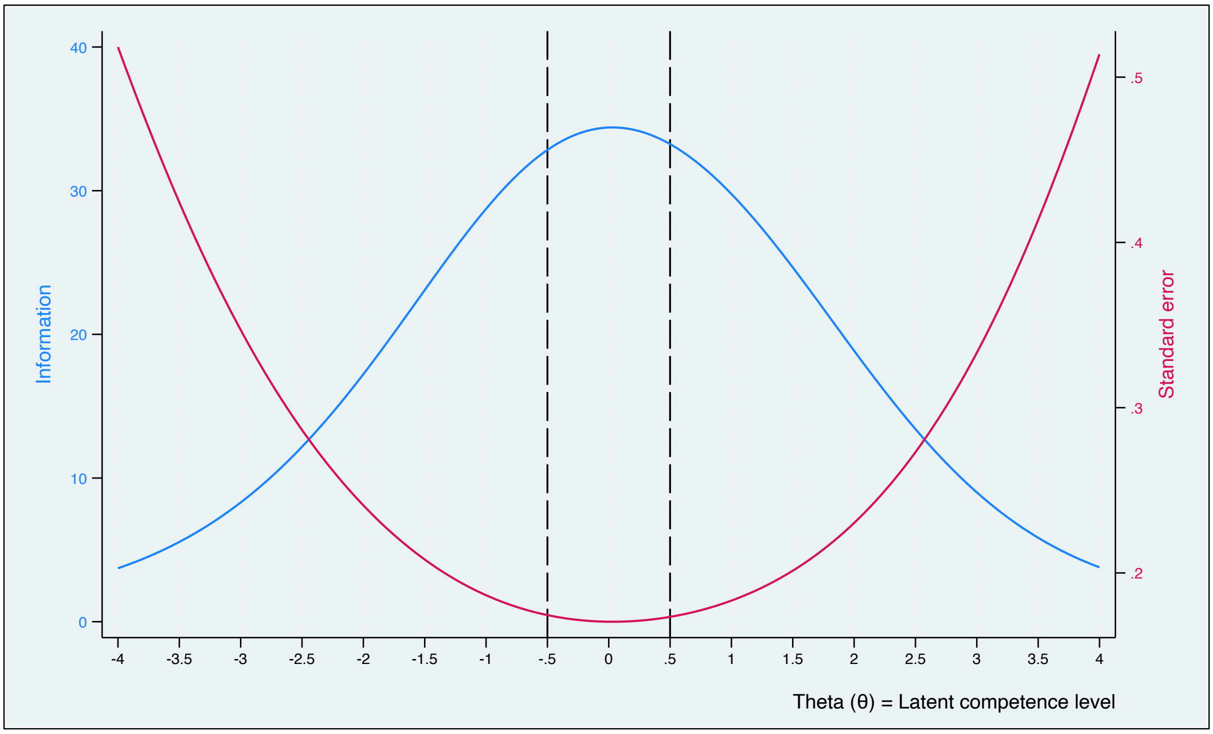
Additional file 5 Figure 1 illustrates the distribution of the empirical Bayes means for the latent trait (θ) which produces posterior distribution using prior information about θ and the likelihood function (actual test scores) by provider category. Overall, this figure corroborates with findings of the one-way ANOVA test (additional file 3) that used total raw scores and proves that θ increases with higher qualification. The pooled θ approximates zero and is normally distributed. Particularly focusing on doctors and midwives (top row), this Figure shows that some doctors and midwives have very low competence while others are very competent and both categories are outliers. Also, the graph of nurses (bottom left) proves that there are fewer nurses who demonstrate having competence comparable to midwives and doctors (e.g., those with $\theta > 1$).

Additional file 6. Determinants of EmONC competence

	Coefficient (95% CI)	S.E. (Bootstrap)	P-value
CEmONC facility (base = BEmONC)*	**0.29 (0.07 – 0.50)**	**0.11**	**0.007**
CEmONC facility (base = BEmONC)**	0.19 (—0.02 – 0.41)	0.11	0.078
CEmONC facility (base = BEmONC)***	0.11 (—0.12 – 0.35)	0.12	0.344
Midwives (base = nurses)	**0.52 (0.20 – 0.83)**	0.16	**0.001**
Doctors (base = nurses)	**1.21 (0.67 – 1.76)**	0.28	**0.000**
Male sex (base = female sex)****	—0.10 (—0.32 – 0.11)	0.11	0.342
Partial time employment (base = full-time)	0.29 (—0.33 – 0.92)	0.32	0.360
EmONC training (base = no training)	**0.36 (0.15 – 0.57)**	0.11	**0.001**
Age in years (continuous)	—0.00 (—0.03 – 0.02)	0.01	0.724
Age in years squared	—0.00 (—0.00 – 0.00)	0.00	0.121
Number of previous deliveries (continuous)	0.00 (—0.00 – 0.00)	0.00	0.639
Number of previous deliveries squared	—0.00 (—0.00 – 0.00)	0.00	0.064

Additional file 6 summarises findings of the linear regression model of EmONC competence score (predicted using the empirical Bayes means) as the dependent variable and facility type, qualification, sex, type of employment, history of EmONC training, age, and the number of deliveries performed before as independent variables. We used 10,000 bootstraps to address heteroskedasticity of residuals and highlighted significant coefficients in bold. *The model includes all providers. **The model includes midwives and nurses. ***The model includes nurses only. ****With the aim to minimise gender and qualification bias (more men are doctors), we excluded doctors from sex variable.

Additional file 7. Test Information Function and validity of the EmONC competence test instrument



Additional file 7 is the Test Information Function (TIF) and illustrates the levels of θ corresponding to the maximum information. Looking at the standard error curve in red plotted against the right scale, we observe that the amount of information provided by the test at θ_i is inversely correlated to the precision with which EmONC competence is estimated at that point. The maximum information (i.e., the maximum precision) lies within the vicinity of θ of zero and remains around 35 between $\theta \pm 0.5$. This instrument is therefore suited for health workers with competence level (θ) such that —0.5 < θ < 0.5.