BMJ Global Health

Women's report of mistreatment during facility-based childbirth: validity and reliability of community survey measures

Hannah Hogan Leslie ,[1,2] Jigyasa Sharma ,[3] Hedieh Mehrtash ,[4] Blair Olivia Berger ,[5] Theresa Azonima Irinyenikan ,[6] Mamadou Dioulde Balde,[7] Nwe Oo Mon,[8] Ernest Maya,[9] Anne-Marie Soumah,[7] Kwame Adu-Bonsaffoh ,[10] Thae Maung Maung ,[8] Meghan A Bohren ,[11] Özge Tunçalp [4]

To cite: Leslie HH, Sharma J, Mehrtash H, *et al*. Women's report of mistreatment during facility-based childbirth: validity and reliability of community survey measures. *BMJ Global Health* 2021;**5**:e004822. doi:10.1136/bmjgh-2020-004822

Handling editor Seye Abimbola

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjgh-2020-004822).

JS and HM contributed equally.

Received 21 December 2020
Accepted 20 July 2021

For numbered affiliations see end of article.

Correspondence to
Dr Hannah Hogan Leslie; hannah.leslie@ucsf.edu

ABSTRACT

Background Accountability for mistreatment during facility-based childbirth requires valid tools to measure and compare birth experiences. We analyse the WHO 'How women are treated during facility-based childbirth' community survey to test whether items mapping the typology of mistreatment function as scales and to create brief item sets to capture mistreatment by domain.

Methods The cross-sectional community survey was conducted at up to 8 weeks post partum among women giving birth at hospitals in Ghana, Guinea, Myanmar and Nigeria. The survey contained items assessing physical abuse, verbal abuse, stigma, failure to meet professional standards, poor rapport with healthcare workers, and health system conditions and constraints. For all domains except stigma, we applied item-response theory to assess item fit and correlation within domain. We tested shortened sets of survey items for sensitivity in detecting mistreatment by domain. Where items show concordance and scale reliability ≥0.60, we assessed convergent validity with dissatisfaction with care and agreement of scale scores between brief and full versions.

Results 2672 women answered over 70 items on mistreatment during childbirth. Reliability exceeded 0.60 in all countries for items on poor rapport with healthcare workers and in three countries for items on failure to meet professional standards; brief scales generally showed high agreement with longer versions and correlation with dissatisfaction. Brief item sets were ≥85% sensitive in detecting mistreatment in each country, over 90% for domains of physical abuse and health system conditions and constraints.

Conclusion Brief scales to measure two domains of mistreatment are largely comparable with longer versions and can be informative for these four distinct settings. Brief item sets efficiently captured prevalence of mistreatment in the five domains analysed; stigma items can be used and adapted in full. Item sets are suitable for confirmation by context and implementation to increase accountability and inform efforts to eliminate mistreatment during childbirth.

Key questions

What is already known?

- Mistreatment during childbirth violates individual rights and may contribute to poor health outcomes for women and people giving birth as well as newborns.
- Instruments for measuring experiences of mistreatment during childbirth have yet to be widely validated and optimised for routine assessment.
- The community survey in the four-country study, 'How women are treated during facility-based childbirth' found frequent but variable experiences of mistreatment across five domains: physical abuse, verbal abuse, failure to meet to professional standards, poor rapport with healthcare providers, health system conditions and constraints.

What are the new findings?

- Secondary analysis of responses from 2672 women provided construct validity evidence in most cases and good item performance for items in each domain.
- Scale reliability was adequate for failure to meet professional standards in three countries and poor rapport with healthcare workers in all study countries. Brief versions of these scales showed strong agreement with full versions.
- Brief sets of survey items were highly sensitive in identifying mistreatment within each of the five domains.

INTRODUCTION

Pregnancy and childbirth are life changing events that should be positive experiences for women, their families and those providing care. This is not possible without provision of quality care that uses a person-centred, rights-based approach to optimise health and well-being for those giving birth and

Key questions

What do the new findings imply?

- Along with the original seven items for assessing stigma, these item sets can be used to identify experiences of mistreatment and monitor these domains of mistreatment within country over time.
- Brief item sets can be used in study settings and tested elsewhere for efficient and sensitive monitoring of women's experiences of domains of mistreatment.
- Comparisons over time and between settings should account for distinct manifestations of women's experiences of mistreatment across contexts and among population subgroups.

their newborns.[1] The WHO recommendations on intrapartum care include guidance on provision of respectful maternity care.[2] They emphasise the fundamental rights of women, newborns and families to equitable access to evidence-based care while recognising the unique needs and preferences of those giving birth and newborns, inclusive of preventing mistreatment during childbirth and promoting respectful care.[2] However, millions of people giving birth in healthcare facilities worldwide are subjected to mistreatment such as physical and verbal abuse, discrimination and neglect.[3] Mistreatment during childbirth is a violation of fundamental rights; it may also negatively impact health outcomes and influence future healthcare seeking behaviour.[4–6] Mistreatment may manifest in different ways across health system contexts and particularly affect women disadvantaged by socioeconomic inequalities,[7] making efforts to define and compare mistreatment more complex.

Reducing mistreatment requires a diagnosis of fundamental drivers of the phenomenon[7]; accountability and evaluation of interventions demand tools to capture the types and prevalence of mistreatment over time and between settings.[8] Individual perspectives are essential to ensuring that health system accountability and improvement efforts centre people's values and preferences for healthcare.[9] However, methodological gaps, including a lack of standardised definitions and instruments as well as considerable variation in choice of population and timing of assessment, have hindered valid and comparable measurement of women's perspectives on mistreatment.[10 11] National and global monitoring of health system performance increasingly recognises the central role of patient experience in measurement, including treatment of women during childbirth.[12] Measurement of respectful and person-centred care for reproductive health is rapidly advancing in many countries.[13–18] A critical question is whether treatment during childbirth can similarly be measured in a valid and comparable way between subgroups in a given health system as well as across health systems.

The four-country WHO study 'How women are treated during childbirth' was designed as a comprehensive, mixed-methods approach to develop and validate tools to measure prevalence of mistreatment of women during childbirth and compare across settings.[19] The first phase built from a systematic review defining a typology of mistreatment including physical abuse, sexual abuse, verbal abuse, stigma and discrimination, failure to meet professional standards of care, poor rapport between women and providers, and health system conditions and constraints.[3] Four study countries—Ghana, Guinea, Myanmar and Nigeria—were purposively sampled to capture a range of health settings and cultures.[20] Primary qualitative work in these settings elicited women's perceptions and experiences of mistreatment[21–23] as well as norms around mistreatment among women and healthcare providers.[24–26] This set of studies identified manifestations of mistreatment in common across settings as well as specific to a single context, such as women reporting health workers whispering as a form of nonverbal insult in Guinea.[22] Items were developed to capture both cross-cutting themes as well as the context-specific insights gathered during formative research.[20] Phase 2 of the study focused on iterative development and testing of two tools to assess the typology of mistreatment—direct observation of labour and birth and a community-based survey—resulting in their fielding in the study countries.[19] Primary analysis focused on prevalence of any mistreatment within domains of the typology, revealing high levels of mistreatment with substantial between-country variation in the specific manifestations.[11] Secondary analysis of the direct observation data from Ghana, Guinea and Nigeria identified consistent measures across these countries for interpersonal abuse, exams and procedures, and unsupportive birth environment.[27] Further use of the community survey tool will be informed by a similar understanding of whether items function as scales to provide domain scores and if subsets of items can provide comparable insight to the original comprehensive item list.

In this analysis, we analyse the community survey tool to test whether items function as scales measuring the domains within the typology of mistreatment, to identify brief item sets that map to the full sets, and to assess comparability of these items sets across the four different health systems and settings in the study. We summarise women's responses according to the hypothesised domains of mistreatment, assess the validity and reliability for each domain within country, test brief versions of each item set against the full set, and describe the potential application of these item sets for comparisons across and within nations.

METHODS

Patient and public involvement

A technical consultation that included representatives from advocacy groups as well as representatives from non-governmental organisations, research organisations, universities, professional associations and United Nations agencies was held in November 2013 and informed the

research questions and design of survey instruments in the WHO study.[28]

Women who recently gave birth in the study countries were involved in content validity testing and providing feedback on the community survey tool prior to data collection. Two group discussions were held with women who recently gave birth in Nigeria to review item clarity, understandability and value. Women recognised value in each item, so all items were retained; items were revised to ensure clarity.[19] Tools were formally piloted in English in Nigeria before being translated by the research team into seven additional languages (Burmese, French, Malinké, Pular, Susu, Twi, Yoruba) and piloted in each site.

Study design and participants

This is a secondary data analysis; procedures for the original study have been described in full previously.[11 19] In brief, 12 hospitals were purposively selected with 3 in each study country (Ghana, Guinea, Myanmar, Nigeria). All facilities were public hospitals in urban settings; number of births per month ranged from 160 to 1506, and staffing types and numbers varied both within and between countries.[11]

Women were eligible for the survey if they were admitted for childbirth at a selected facility, were at least 15 years old, were residents of the facility catchment area (defined for each facility) and were able to and did provide consent. Women were contacted starting 2–3 weeks after birth to schedule the survey; surveys were conducted using digital tablets in a private location and could be conducted up to 8 weeks post partum. Data collection continued until prespecified minimum sample size of 507 in Nigeria (where pilot data had been collected) and 627 per country (209 per facility) in the other countries was met.

Measures

This analysis focused on responses to items within the domains of physical abuse, verbal abuse, stigma and discrimination, failure to meet professional standards of care, poor rapport between women and providers, and health system conditions and constraints from the mistreatment of women during childbirth typology.[3] The most common form of items for this analysis was asking whether a specific form of mistreatment occurred (eg, 'You were shouted or screamed at by a health worker or other staff') and if so, how frequently (eg, once, twice, three or more times, don't know). Some items were asked with Likert-type response options, for instance, 'During my time in hospital for childbirth, I felt ignored by the health workers or staff: Always, most of the time, some of the time, never'. Items regarding professional standards of care referenced a number of possible procedures (eg, caesarean section, episiotomy). If a procedure was received, each woman was asked whether it was explained and whether she agreed to it. Items were coded so that 0 indicated no mistreatment and 1 (binary) or higher values (categorical Likert responses) indicated the presence of mistreatment.

Individual women's characteristics included age in years, language of survey administration, marital status (currently single vs married or cohabitating), education (less than primary vs primary school and above) and primiparity. For convergent validity evidence, we considered women's responses to the item, 'Do you agree or disagree with this statement: Overall, I am satisfied with the services I received during my stay at the hospital for childbirth' and coded level of dissatisfaction from 1=strongly agree to 5=strongly disagree. Satisfaction with care is an outcome of high-quality health systems that is distinct from, but informed by, the experience of care,[9 29] and that may be particularly salient in shaping confidence in and future use of the healthcare system.[30]

Item review

All analysis followed the typology of mistreatment domains.[3] We reviewed previous analysis of these data[11] and assessed response distributions to propose item forms for analysis. The full list of items and frequency of responses is shown by domain in online supplemental table 1. While the primary analysis of the community survey found that 35.4% of women reported experiencing physical or verbal abuse or stigma/discrimination,[11] relatively low numbers of women reported specific subforms of physical and verbal abuse (eg, slapping, pinching, shouting at, insulting). Reports of being shouted at was by far the most common (533 of 2654 women, 20%); reporting a single incident was the most common form of each type of reported abuse. We therefore focused on any occurrence of a type of abuse, rather than frequency. Small numbers of women reported being held down or tied to a bed; we created a composite item of restrained to bed for analysis. Given that <0.5% of the respondents reported 'other' forms of physical and verbal abuse from the defined options, we considered the named types of abuse as comprehensive and eliminated the other item from consideration.

To assess failure to meet professional standards, in keeping with the main study analysis we created an indicator for whether any of four common procedures—vaginal exam, caesarean section, episiotomy and induction of labour—were conducted without informed consent (procedure being both explained to the woman and agreed to). We also excluded an item on skilled attendance during admission due to potentially divergent interpretations among respondents.

In assessing poor rapport between women and providers, the item on interpreter availability when needed was largely not applicable for the populations in this study (1.0% of women needed an interpreter[11]) and was removed for further analysis. Multiple items were asked regarding bed sharing in health system conditions and constraints. Initial review identified differing response patterns by setting; we retained the individual items for further analysis.

Lastly, we excluded the stigma items from subsequent analysis on the basis that these items are not intended to be scaled and are not amenable to reduction without losing essential information; the original items are distinct expressions of forms of stigma against specific groups that merit assessment individually.

In total, 47 items (40 binary, 7 categorical) mapped to the 5 domains for analysis.

Statistical analysis

We primarily used item-response theory (IRT) methods to meet the analytic objectives of identifying whether items performed as a scale in measuring the defined domains of mistreatment and of selecting a subset of items to efficiently and accurately identify women subject to any mistreatment. While comparable in purpose to confirmatory factor analysis (CFA), IRT methods provide three strengths specific to the aims and data of this study: they are intended to confirm the suitability of individual items against a clearly defined construct[31 32] (domain of mistreatment[3]), they enable comparison of item performance in distinct subsets of an overall population,[33] and—in contrast to CFA—they are particularly suitable for binary items.[34] IRT methods have been applied in many areas of clinical and health research,[35] including to validate measures of patient-reported outcomes and consider comparison across settings.[36–39]

Analysis proceeded in three overall steps for each of the domains assessed:

1. Testing full-length item sets within each country to provide evidence of construct validity, to gauge if item sets show sufficient reliability to be considered as a scale, and to assess convergent validity if so.
2. Developing brief item sets that capture any mistreatment with adequate validity and reliability within country.
3. Testing performance of brief scales on pooled data for cross-country comparability.

Methodological details are provided in online supplemental section 2. Briefly, we first limited items for each domain to those that could be assessed within each country and tested model fit using a likelihood ratio test, identified item misfit based on root mean square deviation (RMSD) >0.10, and assessed item concordance by reporting mean expected a posteriori (EAP) scale score for each item response.[40] We assessed differential item functioning (DIF) by sociodemographic characteristics: age, language, marital status, education and parity.[41] DIF indicates variation in responses by a subgroup of respondents conditional on overall scale mean, signalling item misfit for specific respondents that can undermine comparability of scale scores.

We report the EAP reliability for each scale by country, which can be interpreted similarly to Cronbach's alpha: values above 0.60 indicate minimum adequate reliability. For all item sets showing reliability ≥0.60, we tested convergent validity by reporting the unadjusted association of the proposed scales with dissatisfaction with care.

Second, we proposed brief forms of each item set, prioritising capacity to detect any mistreatment and considering test information and scale reliability as applicable. We assessed sensitivity by comparing report of mistreatment based on the brief item set to women reporting mistreatment on any of the original items by domain (including a response of neutral, agree or strongly agree for categorical items). Where items could be summarised into scales, we quantified the agreement of the brief and full scales in classifying women's experiences of mistreatment by categorising women into quintiles on each scale and calculating a weighted kappa statistic.

Finally, we assessed the performance of brief scales in enabling comparisons between countries by repeating the model and item analysis on a pooled sample of all respondents and testing DIF by country. Analyses were conducted in R V.3.5.2 (R Foundation for Statistical Computing) with packages TAM and psychotree[41 42] and in Stata (StataCorp. 2019. Stata Statistical Software: Release 16. College Station, TX: StataCorp LLC).

RESULTS

Respondents

Two thousand six hundred seventy-two women were included in this analysis; table 1 describes the study sample. Most women were married or cohabitating and over half were primiparous (from 44% of respondents in Nigeria to 66% of respondents in Ghana). Between 8% (Guinea) and 15% (Nigeria) of respondents expressed a lack of satisfaction with care (neutral or disagreed). Column 2 in table 2 lists the items considered for statistical analysis; results are presented for each domain below.

Physical abuse

Full item set

The most common forms of physical abuse were application of forceful downward pressure on the abdomen (6% overall, up to 16% in Guinea) followed by being slapped (4% overall, up to 11% in Nigeria, online supplemental figure S1). We removed four items with no reports of mistreatment in at least one country: being kicked, punched, hit or gagged. Remaining items were modelled using a one-parameter logistic (1PL, Rasch) model for respondents in Ghana and Myanmar; the two-parameter logistic (2PL) model showed better fit to responses from Guinea and Nigeria. Item responses were correlated to scale scores except for report of being pinched among respondents in Guinea (online supplemental table S2); all items demonstrated good fit in all countries and no DIF within country. Reliability of the proposed scale was poor (0.05 in Ghana to 0.24 in Nigeria, table 3); selected items were considered as an item set for subsequent analysis.

Brief item set

We did not shorten the four-item physical abuse set given its already brief nature; the four item-set was highly

Table 1 Study sample

	Ghana (N=836)	Guinea (N=644)	Myanmar (N=631)	Nigeria (N=561)	Total (N=2672)
Age					
Median (Q1, Q3)	29.0 (25.0, 33.0)	23.0 (19.0, 28.0)	27.0 (23.0, 32.0)	30.0 (26.0, 34.0)	28.0 (23.0, 32.0)
Language					
English	177 (21%)	0 (0%)	0 (0%)	233 (42%)	412 (15%)
French	0 (0%)	145 (23%)	0 (0%)	0 (0%)	145 (5%)
Pular	0 (0%)	100 (16%)	0 (0%)	0 (0%)	100 (4%)
Malinké	0 (0%)	69 (11%)	0 (0%)	0 (0%)	69 (3%)
Susu	0 (0%)	330 (51%)	0 (0%)	0 (0%)	330 (12%)
Twi	659 (79%)	0 (0%)	0 (0%)	0 (0%)	659 (25%)
Yoruba	0 (0%)	0 (0%)	0 (0%)	328 (58%)	328 (12%)
Burmese	0 (0%)	0 (0%)	629 (>99%)	0 (0%)	629 (24%)
Unknown	0 (0%)	0 (0%)	2 (<1%)	0 (0%)	2 (<1%)
Marital status					
Single	134 (16%)	45 (7%)	18 (3%)	34 (6%)	231 (9%)
Married/cohabitating	700 (84%)	599 (93%)	613 (97%)	527 (94%)	2439 (91%)
Unknown	2 (<1%)	0 (0%)	0 (0%)	0 (0%)	2 (<1%)
Education					
Less than primary	453 (54%)	555 (86%)	303 (48%)	57 (10%)	1368 (51%)
Completed primary school or above	383 (46%)	89 (14%)	328 (52%)	504 (90%)	1304 (49%)
Parity					
Multiparous	282 (34%)	233 (36%)	275 (44%)	316 (56%)	1106 (41%)
Zero previous births	550 (66%)	411 (64%)	355 (56%)	244 (44%)	1560 (59%)
Unknown	4 (<1%)	0 (0%)	1 (<1%)	1 (<1%)	6 (<1%)
Satisfied with services overall					
Strongly agree	284 (34%)	354 (55%)	131 (21%)	149 (27%)	918 (34%)
Agree	456 (55%)	231 (36%)	428 (68%)	328 (58%)	1443 (54%)
Neutral	50 (6%)	16 (2%)	3 (0%)	31 (6%)	100 (4%)
Disagree	31 (4%)	28 (4%)	55 (9%)	36 (6%)	150 (6%)
Strongly disagree	10 (1%)	14 (2%)	14 (2%)	17 (3%)	55 (2%)
Unknown	5 (<1%)	1 (<1%)	0 (0%)	0 (0%)	6 (<1%)

sensitive for any reported physical mistreatment (92% in Guinea to 98% in Ghana, figure 1).

Verbal abuse

Full item set

One item was removed from the verbal abuse scale due to 0% prevalence within a country sample (negative comments about the baby's appearance). The most common forms of verbal abuse were being shouted at (20%), scolded (10%) and threatened with a poor outcome (7%). Likelihood ratio tests rejected the 1PL model in favour of the 2PL in all samples except Myanmar, where prevalence of verbal abuse items was notably lower. Item responses indicative of mistreatment were linked to higher scale scores in all cases (online supplemental table S3A); no items exceeded the threshold for misfit. DIF analysis identified differential functioning by age among respondents in Ghana. We found that the 10-item scale shown in table 2, column 3 performed well, with no within-country DIF, higher scale scores by report of mistreatment for each item (online supplemental table S3B), and good item fit. Reliability of this scale was low, ranging from 0.35 among respondents in Myanmar to barely adequate at 0.61 in the Nigerian sample (table 3). The scale was associated with dissatisfaction among women in Nigeria (table 4).

Brief item set

Four items covered distinct content and captured the most commonly reported forms of verbal abuse in each setting: being shouted at, scolded, threatened with medical procedure, threatened with poor outcome (last

Table 2 Items considered for analysis and items included in final item sets

		Final item sets	
	Items for initial analysis based on original community survey tool	**Full length**	**Brief**
Physical abuse	Pinched	X	X
	Kicked		
	Slapped	X	X
	Punched		
	Hit		
	Gagged		
	Restrained to bed	X	X
	Forceful downward pressure on abdomen	X	X
Verbal abuse	Shouted at	X	X
	Insulted	X	
	Scolded	X	X
	Mocked	X	
	Hissed at	X	
	Negative comments about woman's physical appearance *		
	Negative comments about baby's physical appearance †		
	Negative comments about woman's sexual activity	X	X
	Threatened with a medical procedure	X	
	Threatened with physical violence		
	Threatened with poor outcome	X	X
	Threatened with withheld care	X	
	Blamed for something that happened	X	
Failure to meet professional standards of care	Lack of informed consent (any of four procedures)	X	X
	During vaginal exam, private information shared so others could hear	X	
	Vaginal exam conducted in a way that other people could see	X	X
	Painful vaginal exams	X	
	Pain relief not provided appropriately ‡	X	X
	Ignored by health workers§	X	X
	Neglected by health workers§	X	
	Felt a nuisance §	X	
	Waited long periods §	X	X
	Skilled birth attendant absent when baby born	X	X

Continued

Table 2 Continued

	Items for initial analysis based on original community survey tool	Final item sets	
		Full length	Brief
Poor rapport with healthcare workers	Healthcare worker not responsive to questions or concerns §		
	Lack of emotional support §	X	X
	Healthcare worker did not listen to concerns §	X	X
	Birth companion not allowed	X	X
	Lacked access to water or other fluids	X	
	Not allowed to eat	X	
	Not told could move during labour	X	X
	Not allowed to give birth in preferred position	X	
	Detained due to inability to pay bills		
Health system conditions and constraints	Lack of privacy/curtains	X	X
	No bed to self during labour	X	
	No bed to self during childbirth	X	
	No bed to self post partum	X	X¶
	Shared a bed at any time	X	X
	Asked for a bribe	X	X
	Made to clean up after oneself	X	

*Weight, private parts, cleanliness or other parts of the woman's body
†Appearance, sex or other aspects
‡Not offered, requested and not received, denied
§Categorical response options
¶Included for use in Ghana

Table 3 Reliability of scales by country

		EAP reliability		Weighted kappa statistic brief versus full-length scale
		Full-length scale	Brief scale	
Physical abuse	Ghana	0.05	NA	NA
	Guinea	0.11	NA	NA
	Myanmar	0.08	NA	NA
	Nigeria	0.24	NA	NA
Verbal abuse	Ghana	0.50	0.43	NA
	Guinea	0.47	0.41	NA
	Myanmar	0.35	0.29	NA
	Nigeria	0.61	0.52	0.90 (0.87, 0.92)
Failure to meet professional standards of care	Ghana	0.85	0.69	0.68 (0.64, 0.72)
	Guinea	0.58	0.48	NA
	Myanmar	0.79	0.66	0.88 (0.85, 0.91)
	Nigeria	0.85	0.67	0.84 (0.80, 0.87)
Poor rapport with healthcare workers	Ghana	0.74	0.74	0.95 (0.94, 0.96)
	Guinea	0.79	0.79	0.97 (0.96, 0.97)
	Myanmar	0.67	0.69	0.88 (0.86, 0.91)
	Nigeria	0.77	0.78	0.89 (0.87, 0.91)
Health system conditions and constraints	Ghana	0.64	0.59	0.95 (0.93, 0.97)
	Guinea	0.52	0.48	NA
	Myanmar	0.43	0.46	NA
	Nigeria	0.24	0.34	NA

EAP, expected a posteriori; NA, not applicable.

column, table 2). Responses indicating mistreatment were linked to higher scale scores (online supplemental table S3C), but reliability was below 0.60 in all countries.

Sensitivity of the four items for detecting any verbal abuse ranged from 86% in Nigeria to 92% in Guinea (figure 1).

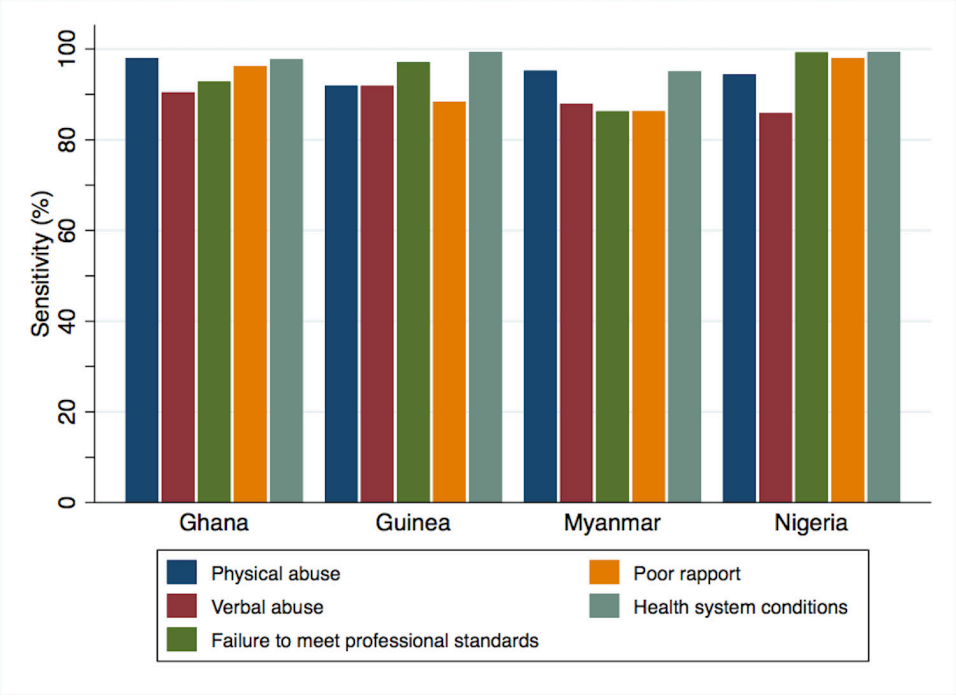

Figure 1 Sensitivity of brief item sets for detecting any mistreatment by domain.

Table 4 Convergent validity evidence—association of scales with dissatisfaction with care received, linear regression models

		Association with level of dissatisfaction Beta	
		Full-length scale	Brief scale
Verbal abuse	Nigeria	0.39*	NA
Failure to meet professional standards of care	Ghana	0.42*	0.31*
	Myanmar	0.42*	0.46*
	Nigeria	0.38*	0.45*
Poor rapport with healthcare workers	Ghana	0.41*	0.40*
	Guinea	0.47*	0.47*
	Myanmar	0.38*	0.39*
	Nigeria	0.44*	0.44*
Health system conditions and constraints	Ghana	0.16*	0.15*

*P<0.05

Failure to meet professional standards of care

Full item set

The 10-item set (table 2) for failure to meet professional standards of care included four categorical items. Women frequently reported lack of informed consent (56% had at least one of the four procedures without fully informed consent) and painful vaginal exams (50% overall, from 6% in Myanmar to 73% in Ghana), while experiences such as absence of a skilled attendant when baby was born was quite rare (2% across all respondents) (online supplemental figure S4). The 2PL model improved fit for all country samples. All items met the threshold for good item fit (RMSD <0.10). Domain scale scores were generally higher for individual item responses indicating mistreatment, although not for all response steps in categorical items and not consistently for items on visual and aural privacy or in Myanmar for the item on attendant at birth (online supplemental table S6A).

A number of items showed DIF assessment by demographic subgroup relative to the rest of each scale, by primiparity in Ghana and by language in Guinea (four languages) and Nigeria (two languages) (online supplemental figure S5). Removing selected items did not change these results; we proceeded with the full item set for initial scales. As shown in table 3, reliability of the 10-item scale was inadequate (0.58) among respondents in Guinea and good (0.79–0.85) among respondents in the other country samples. Scale scores were significantly associated with dissatisfaction with care in the three samples assessed (table 4).

Brief item set

Six items provided information across the range of respondents (table 2, column 4). All items showed good overall fit; the relationship between item-specific report of mistreatment and domain score was weakest for categorical item responses and for respondents in Ghana (online supplemental table S6B). DIF was still present in Ghana and Guinea. Reliability of the six-item scale did not meet the threshold of 0.60 among respondents in Guinea (0.48) (table 3). Weighted kappa statistics in the other three settings support the brief scale in capturing much of the information of the full-length scale for identifying women experiencing more mistreatment, ranging from 0.68 in Ghana to 0.88 in Myanmar. The brief scale was associated with dissatisfaction in each of these three country samples (table 4). As shown in figure 1, sensitivity of the six items to any mistreatment in this domain was very high in Ghana (93%), Guinea (97%) and Nigeria (99%), and moderate in Myanmar (86%).

Poor rapport between women and healthcare providers

Full item set

The item pool for poor rapport included three categorical items; overall approximately 30% of women reported some level of mistreatment regarding providers' listening, being responsive and providing emotional support, with higher levels of these types of mistreatment in Myanmar (online supplemental figure S6). In contrast, very few women reported not being allowed a birth companion in Myanmar (<1%) compared with more than half of women reporting this Ghana, Guinea and Nigeria. The 2PL model improved fit for all country samples. Domain scores increased with response options for categorical items except for the highest categories among respondents in Guinea; binary items such as lack of access to water and not being allowed to deliver in a preferred position showed inconsistent links to domain scores across countries (online supplemental table S7A). Items demonstrated good overall fit, but several showed DIF by demographic subgroup (online supplemental figure S7). We removed lack of responsiveness to questions or concerns and being detained due to inability to pay bills and refit the seven items shown in table 2, column 3. The resulting scale showed no DIF in Myanmar or Nigeria, though responses in Ghana and Guinea differed based on language of the survey and primiparity conditional on responses to other items. Reliability of the seven-item scale ranged from 0.67 among respondents in Myanmar to 0.79 among respondents in Guinea (table 3). Convergent validity was supported by significant associations between scale scores and dissatisfaction with care (table 4).

Brief item set

Four items composed the brief scale: lack of emotional support, not listening to concerns, birth companion not allowed and not told she could move during labour (table 2, column 4). Item-specific responses indicating mistreatment were linked to higher scale scores except for 'disagree strongly' options in Guinea (online supplemental table S7C). All items showed good overall fit, and DIF assessment was comparable to the full scale. Reliability of the four-item scale was as good or better than

the full-length scale (table 3). Weighted kappa statistics ranged from 0.88 in Myanmar to 0.97 in Guinea, indicating strong agreement. The brief scale was associated with dissatisfaction in each country sample (table 4). As shown in figure 1, these items were highly sensitive to any mistreatment on this domain among respondents in Ghana (96%) and Nigeria (98%) and slightly less so among respondents in Guinea (88%) and Myanmar (86%).

Health system constraints

Full item set

Responses on the seven items on health system conditions and constraints differed among women in Ghana compared with the other three countries (online supplemental figure S8). For respondents in Guinea, Myanmar and Nigeria, the items on lack of privacy or curtains and being asked for a bribe were the two main forms of mistreatment. One third of women in Ghana answered no to questions on having a bed to oneself during childbirth and post partum compared with <8% in all other countries, although these responses were not reflected in the single item on sharing a bed at any time or the item on lack of privacy. Having to clean up after oneself was reported only in Myanmar (16% compared with <1% in other study countries).

2PL models improved fit in all countries. Item-specific report of mistreatment was linked to higher domain scores in Guinea and Nigeria but not for one item in Myanmar; only the items on not having a bed to oneself showed concordance among respondents in Ghana (online supplemental table S8A). All items passed the threshold for item fit. Tests for DIF identified differences by sociodemographic group, mainly primiparity and language of survey response, though items affected differed by country. Reliability was adequate among respondents in Ghana (0.64), table 3. Given the poor concordance of item responses and scale scores among respondents in Ghana, we focus on the use of items as an item set in all countries.

Brief item set

Items on lack of privacy and requests for a bribe encompassed the majority of mistreatment in this domain among respondents in Guinea, Myanmar and Nigeria. Including the item on bed share at any time for all settings and adding the item on no bed to oneself post partum for women in Ghana resulted in an item set with high sensitivity to all forms of mistreatment in this domain (95% in Myanmar to 99% in Guinea and Nigeria, figure 1).

Cross-national comparisons

Analysis supported scales capturing failure to meet professional standards of care (three countries) and poor rapport with healthcare workers (all countries). We assessed the proposed brief scales for comparability across countries. Both scales demonstrated DIF between countries. Comparing the model parameters for the pooled sample and for each country (online supplemental table S9) shows that item discrimination and difficulty varied between countries in magnitude and in ordering within scales, making it difficult to quantify the degree of mistreatment across countries with these scales.

DISCUSSION

We conducted a secondary analysis of over 2600 women's experiences in childbirth across four country settings to test full and brief item sets to address five domains in the typology of mistreatment: physical and verbal abuse, failure to meet professional standards of care, poor rapport with healthcare workers, and health system conditions and constraints. Reliability was adequate to treat item sets as a scale producing a summary score among respondents in three study sites for failure to meet professional standards of care and in all sites for poor rapport with healthcare workers. These scales were associated with dissatisfaction with care in each setting, and brief scales classified women's experience of mistreatment similarly to full-length scales. Evidence of mistreatment on brief item sets standardised across countries was generally a sensitive indicator of any mistreatment for each domain. Based on this evidence from urban hospitals in four countries, brief item sets can provide an efficient and sensitive method of identifying women experiencing these domain of mistreatment during childbirth.

Items within the domains of failure to meet professional standards and poor rapport with healthcare workers demonstrated the reliability and consistency to use as scales in most study settings. Lower concordance of categorical item responses with overall scale scores suggests that greater mistreatment on categorical items may not always co-occur with other types of mistreatment and/or that categorical response options may be understood differently by respondents. Evidence of DIF by survey language, particularly in Guinea where the survey was conducted in four languages, could also reflect some divergence in how respondents interpreted categorical response items. DIF by characteristics such as parity may reflect distinct expectations for those with prior experience of the birthing process. Population prevalence of any mistreatment for these domains can be compared directly based on item sets, while scale scores should be calculated by strata to avoid bias due to group composition when units such as facilities are compared. Evidence on reliability, validity and sensitivity of brief item sets for these domains suggests that they can be used to identify any mistreatment in all study settings, and as scales to quantify degree of mistreatment within each country except for failure to meet professional standards in Guinea. Scale scores are not directly comparable across study countries.

Items within domains of physical abuse, verbal abuse and health system conditions and constraints are better used as item sets than scales intended to distinguish across a spectrum of mistreatment. Brief item sets were

over 85% (verbal abuse) and 90% (physical abuse and health system conditions and constraints) sensitive. The item set for health system conditions and constraints differed across countries, with one item added to better reflect responses in Ghana. Although formative research supported not having one's own bed as a form of mistreatment, the frequent report of this practice in Ghana did not concord with responses on privacy and bed sharing; it is possible that women's responses may reflect factors such as facility practice of moving postpartum women from the labour ward to a different ward to make way for other labouring women. Inclusion of items on having one's own bed for the item set in Ghana warrant further consideration in this setting.

Across all domains, the finding that 3–6 items per domain provided high but not perfect sensitivity in all cases underscores that a single item per domain will not be a reliable proxy for level of mistreatment within or between settings. This finding is not entirely unexpected, as the detailed qualitative work in each country identified country-specific manifestations of types of abuse, such as forceful downward pressure on the abdomen in Guinea[22] and slapping as a way of improving the birth outcome in Ghana and Nigeria.[21 23]

This analysis removed the item on need for an interpreter, which may be salient in specific settings. Notably, we did not consider items on stigma as amenable to scaling or reduction. Stigma and discrimination are critical elements of mistreatment and poor experiences of healthcare; we suggest that future research and programming consider the seven stigma and discrimination items from our original tools, and adapt (as needed) for the context of interest. Measuring stigma and discrimination is essential to assess health equity and to ensure that no one is left behind.

Results of this study can be compared with development of related tools on respectful and patient-centred maternity care in other countries, which share common content around respect and communication with women as well as stigma and discrimination.[13 16 18 43] Specific item decisions are directly comparable in some cases: items on wait time, visual privacy, labour companion and healthcare workers paying attention when help needed were included in the brief item sets in this study and the person-centred maternity care (PCMC) scale validated in Kenya. Items including access to food and water and aural privacy were removed in both cases.[13] Use of the PCMC scale in Kenya, India and Ghana provided evidence of adequate reliability for overall scale creation,[14] as did this assessment on failure to meet professional standards of care and poor rapport with healthcare workers, the most similar domains to the PCMC in terms of content and response types. The measures diverge in items on physical abuse, verbal abuse and informed consent for procedures, which are asked each in a single item in the PCMC scale but elicited based on specific types of abuse or procedures in this study. Similarly, items on stigma are asked separately by attribute discriminated against (eg, age, HIV status, religion) in this study, but often as a combined item or items in other scales.[13 16 18] This focus on individual forms of mistreatment contributed to the recommendation of item sets rather than scales for comprehensive measurement of abuse and stigma. Assessments of disrespect and abuse commonly use multiple items to elicit more complete and specific responses than obtained using composite items.[10 44 45] This study combined with existing findings confirms that core constructs of mistreatment can be measured in multiple settings using individual self-report; the brief item sets tested here span mistreatment domains and demonstrate high sensitivity in detecting mistreatment, making them well suited to comprehensive detection of mistreatment. Use of the brief item sets and other scales in the same population would be needed to compare their performance directly.

Findings are limited in several ways. Study facilities were high-volume public facilities in urban areas[11]; if the patterns and types of mistreatment are distinctive in such facilities, the findings may not generalise. Evidence from settings other than the study countries does suggest variability in level and in some cases type of mistreatment by facility characteristics.[46–48] Further assessment in smaller facilities and in rural areas is warranted. The IRT analysis assumes independent item response conditional on the latent trait and unidimensionality of each domain. Violations of these assumptions would invalidate results on model fit and reliability. We consider dissatisfaction with care as an external criterion to support scale validity. The assumption that patients translate negative experiences into a dissatisfaction rests on expectations of care—assessing experiences against what's feasible and expected—and attribution of responsibility to providers[49]; for instance, patients with negative experiences related to health system conditions and constraints may report satisfaction relative to their expectations and to what providers are responsible for. Qualitative work with women in Nigeria, Guinea and Myanmar during the formative phase of the WHO study suggested that women found most types of mistreatment unacceptable,[24–26] but that perceived justifications for mistreatment such as aiding in labour could help to shape ratings of satisfaction or dissatisfaction. A secondary analysis using data from this study found that women who reported mistreatment were more likely to report lower satisfaction with care.[29] Use of alternative measures of mistreatment would provide further validation in future studies.

This analysis builds on the strengths of the WHO 'How women are treated during facility-based childbirth' study, which developed tools specifically to capture mistreatment based on extensive formative research and pretesting in four settings and tested them at scale. The use of a small number of facilities in the sample should reduce variability in the underlying construct. Surveys were carried out in the weeks after birth to bolster recall and in the community to reduce social desirability bias from exit interviews; items addressed a wide range of

manifestations of mistreatment to capture women's experiences as broadly as possible.

This work has a number of implications for research. Measuring women's perspectives is inherently complex due to changing expectations and perceptions of health services.[50 51] The analysis of labour observations from the same study found evidence for cross-country comparability in items and scores for a scale on interpersonal abuse and item sets for exams and procedures and unsupportive birth environment, potentially reflecting the comparability of trained observers applying prestandardised definitions to the widely varying experience of childbirth.[27] Woman-centred measures of quality of care and birth experiences are critical to evaluating maternity care, as women are the best experts on their own experiences,[9] but such assessments must consider the power imbalance that contributes to mistreatment and may shape the perception and reporting of it.[7] As a priority for future research, triangulation between the observations of care and women's self-reports will help to identify what types of mistreatment can be monitored with greater sensitivity using direct observation, particularly for marginalised women. Further assessment of the expectations for childbirth care and the factors that shape them is also warranted.[52 53] Finally, efforts to identify and monitor mistreatment would be facilitated by research quantifying the number of respondents and the minimum sufficient number of labour observations required to reliably assess communities and facilities.

In considering ongoing measurement for monitoring and spurring action, the original community survey instrument provides a comprehensive assessment of all domains and can be summarised by individual item. Brief item sets proposed here provide shorter but generally highly sensitive means of identifying mistreatment by domain in the distinct study settings of hospitals in urban Ghana, Guinea, Myanmar and Nigeria. Full-length and brief scales support synthesis of two mistreatment domains that can be monitored and reported within country over time and that classify women's experience of mistreatment similarly. Measurement of stigma was not subject to assessment, but should be included based on the original seven items.

This analysis as well as other in depth analyses of the study findings have identified substantial differences in how mistreatment is experienced in distinct healthcare settings and how forms of mistreatment may be linked.[54] As a whole, this body of work confirms that mistreatment is complex and cannot be measured by only one or two items standardised across populations and health system settings. Interventions to reduce mistreatment will require context-specific understanding of mechanisms and drivers within the health system. These item sets provide a means of community-based assessment to identify mistreatment domains and hold the health system accountable; they can be incorporated into ongoing efforts such as Demographic and Health Surveys and more targeted surveys intended to inform and ignite action for improvement. Efficient and sensitive assessment of the domains of mistreatment can demand accountability and compel action towards the ultimate goal of eliminating mistreatment and improving quality of care for women and people giving birth across the world.

Author affiliations
[1]Global Health and Population, Harvard University T H Chan School of Public Health, Boston, Massachusetts, USA
[2]Division of Prevention Science, University of California San Francisco, San Francisco, California, USA
[3]Chief Economist's Office, Human Development Group, World Bank Group, Washington, District of Columbia, USA
[4]Department of Sexual and Reproductive Health and Research, including UNDP/UNFPA/UNICEF/WHO/World Bank Special Programme of Research, Development and Research Training in Human Reproduction (HRP), World Health Organization, Geneve, Switzerland
[5]Population, Family and Reproductive Health, Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland, USA
[6]Department of Obstetrics and Gynaecology, University of Medical Sciences Teaching Hospital Complex, Akure, Ondo State, Nigeria
[7]Cellulle de Recherche en Sante de la Reproduction en Guinee (CERREGUI), University National Hospital-Donka, Conakry, Guinea
[8]Department of Medical Research, Ministry of Health and Sports, Yangon, Myanmar
[9]School of Public Health, University of Ghana, Accra, Ghana
[10]Department of Obstetrics and Gynecology, University of Ghana Medical School, Accra, Ghana
[11]Gender and Women's Health Unit, Centre for Health Equity, University of Melbourne School of Population and Global Health, Melbourne, Victoria, Australia

Twitter Hedieh Mehrtash @hediehmm, Blair Olivia Berger @blair_berger and Özge Tunçalp @otuncalp

Acknowledgements The authors thank Soe Soe Thwin for review of the analysis plan and Olusoji Adeyanju, Richard Adanu, Boubacar Diallo, Alpha Oumar Sall, and Joshua Vogel for their contributions to the conduct and analysis of the primary analysis. We would like to express our sincere gratitude to the women and providers who participated in this study. We are thankful to the research team in Guinea, Ghana, Nigeria and Myanmar, for their great effort and excellent work provided to this project which would not have been possible without their contribution.

Contributors Conceptualised this analysis: HHL, JS, HM, MAB, ÖT. Conducted training, data collection, data management: MAB, HM, TAI, MDB, TMM, NOM, EM, A-MS, KA-B. Methodology: HHL, JS, HM, BOB, ÖT. Formal analysis and original draft writing: HHL. Supervision: MAB, ÖT. All authors involved in data interpretation and review of the final manuscript.

Funding This research was funded by the support of the American People through the United States Agency for International Development (USAID) and the UNDP/UNFPA/UNICEF/WHO/World Bank Special Programme of Research, Development and Research Training in Human Reproduction (HRP), Department of Sexual and Reproductive Health and Research, WHO.

Competing interests HHL declares research support from the Bill & Melinda Gates Foundation, the World Bank and ICF International outside the scope of this work.

Patient consent for publication Not required.

Ethics approval This secondary analysis was declared not human subjects research by the Institutional Review Board at the Harvard TH Chan School of Public Health (IRB18-1392). The original study was approved by the WHO Ethical Review Committee (protocol: A65880) and the WHO Human Reproduction Programme (HRP) Review Panel on Research Projects, and in-country ethical committees; Le Comité National d'Ethique pour la Recherche en Santé (Guinea); Federal Capital Territory Health Research Ethics Committee (Nigeria); Research Ethical Review Committee, Oyo State (Nigeria); State Health Research Ethics Committee of Ondo State (Nigeria); Ethical Review Committee of the Ghana Health Service (Ghana); Ethical and Protocol Review Committee of the College of Health Sciences, University of Ghana (Ghana); and Ethics Review Committee, Department of Medical Research (Myanmar).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon request. The analytic study dataset from the "WHO Study: How women are treated during facility-based childbirth" is de-identified and archived through WHO/HRP's electronic record management system. Data requests with an expression of interest in pursuing multi-country secondary analyses with a specific research question can be made to srhmph@who.int. More information about the study tools are available here: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-018-0603-x and the primary publication from the study here: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(19)31992-0/fulltext.

ORCID iDs
Hannah Hogan Leslie http://orcid.org/0000-0002-7464-3645
Jigyasa Sharma http://orcid.org/0000-0001-8860-0710
Hedieh Mehrtash http://orcid.org/0000-0003-4991-616X
Blair Olivia Berger http://orcid.org/0000-0002-7962-0522
Theresa Azonima Irinyenikan http://orcid.org/0000-0002-5594-037X
Kwame Adu-Bonsaffoh http://orcid.org/0000-0002-3741-6646
Thae Maung Maung http://orcid.org/0000-0002-1265-3813
Meghan A Bohren http://orcid.org/0000-0002-4179-4682
Özge Tunçalp http://orcid.org/0000-0002-5370-682X

REFERENCES

1 Shakibazadeh E, Namadian M, Bohren MA, *et al*. Respectful care during childbirth in health facilities globally: a qualitative evidence synthesis. *BJOG* 2018;125:932–42.
2 World Health Organization. *WHO recommendations on intrapartum care for a positive childbirth experience*. Geneva, Switzerland: World Health Organization, 2018.
3 Bohren MA, Vogel JP, Hunter EC, *et al*. The mistreatment of women during childbirth in health facilities globally: a mixed-methods systematic review. *PLoS Med* 2015;12:e1001847.
4 Bohren MA, Hunter EC, Munthe-Kaas HM, *et al*. Facilitators and barriers to facility-based delivery in low- and middle-income countries: a qualitative evidence synthesis. *Reprod Health* 2014;11:71.
5 Khosla R, Zampas C, Vogel JP, *et al*. International human rights and the mistreatment of women during childbirth. *Health Hum Rights* 2016;18:131–43.
6 Zampas C, Amin A, O'Hanlon L, *et al*. Operationalizing a human Rights-Based approach to address mistreatment against women during childbirth. *Health Hum Rights* 2020;22:251–64.
7 Sen G, Reddy B, Iyer A. Beyond measurement: the drivers of disrespect and abuse in obstetric care. *Reprod Health Matters* 2018;26:6–18.
8 Afulani PA, Buback L, McNally B, *et al*. A rapid review of available evidence to inform indicators for routine monitoring and evaluation of Respectful maternity care. *Glob Health Sci Pract* 2020;8:125–35.
9 Larson E, Sharma J, Bohren MA, *et al*. When the patient is the expert: measuring patient experience and satisfaction with care. *Bull World Health Organ* 2019;97:563–9.
10 Sando D, Abuya T, Asefa A, *et al*. Methods used in prevalence studies of disrespect and abuse during facility based childbirth: lessons learned. *Reprod Health* 2017;14:127.
11 Bohren MA, Mehrtash H, Fawole B, *et al*. How women are treated during facility-based childbirth in four countries: a cross-sectional study with labour observations and community-based surveys. *Lancet* 2019;394:1750–63.
12 Quality of care network - resources and data related to maternal, newborn and child health quality of care measurement [Internet]. Available: https://www.who.int/data/maternal-newborn-child-adolescent/documents/mca [Accessed 25 Aug 2020].
13 Afulani PA, Diamond-Smith N, Golub G, *et al*. Development of a tool to measure person-centered maternity care in developing settings: validation in a rural and urban Kenyan population. *Reprod Health* 2017;14:118.
14 Afulani PA, Phillips B, Aborigo RA, *et al*. Person-centred maternity care in low-income and middle-income countries: analysis of data from Kenya, Ghana, and India. *Lancet Glob Health* 2019;7:e96–109.
15 Sudhinaraset M, Afulani PA, Diamond-Smith N, *et al*. Development of a Person-Centered family planning scale in India and Kenya. *Stud Fam Plann* 2018;49:237–58.
16 Sheferaw ED, Mengesha TZ, Wase SB. Development of a tool to measure women's perception of respectful maternity care in public health facilities. *BMC Pregnancy Childbirth* 2016;16:67.
17 Gurung R, Ruysen H, Sunny AK, *et al*. Respectful maternal and newborn care: measurement in one EN-BIRTH study hospital in Nepal. *BMC Pregnancy Childbirth* 2021;21:228.
18 Vedam S, Stoll K, Rubashkin N, *et al*. The mothers on respect (MOR) index: measuring quality, safety, and human rights in childbirth. *SSM Popul Health* 2017;3:201–10.
19 Bohren MA, Vogel JP, Fawole B, *et al*. Methodological development of tools to measure how women are treated during facility-based childbirth in four countries: labor observation and community survey. *BMC Med Res Methodol* 2018;18:132.
20 Vogel JP, Bohren MA, Tunçalp Ö, *et al*. How women are treated during facility-based childbirth: development and validation of measurement tools in four countries - phase 1 formative research study protocol. *Reprod Health* 2015;12:60.
21 Maya ET, Adu-Bonsaffoh K, Dako-Gyeke P, *et al*. Women's perspectives of mistreatment during childbirth at health facilities in Ghana: findings from a qualitative study. *Reprod Health Matters* 2018;26:70–87.
22 Balde MD, Diallo BA, Bangoura A, *et al*. Perceptions and experiences of the mistreatment of women during childbirth in health facilities in guinea: a qualitative study with women and service providers. *Reprod Health* 2017;14:3.
23 Bohren MA, Vogel JP, Tunçalp Ö, *et al*. Mistreatment of women during childbirth in Abuja, Nigeria: a qualitative study on perceptions and experiences of women and healthcare providers. *Reprod Health* 2017;14:9.
24 Balde MD, Bangoura A, Diallo BA, *et al*. A qualitative study of women's and health providers' attitudes and acceptability of mistreatment during childbirth in health facilities in guinea. *Reprod Health* 2017;14:4.
25 Bohren MA, Vogel JP, Tunçalp Ö, *et al*. "*By slapping their laps, the patient will know that you truly care for her*": A qualitative study on social norms and acceptability of the mistreatment of women during childbirth in Abuja, Nigeria. *SSM Popul Health* 2016;2:640–55.
26 Maung TM, Show KL, Mon NO, *et al*. A qualitative study on acceptability of the mistreatment of women during childbirth in Myanmar. *Reprod Health* 2020;17:56.
27 Berger BO, Strobino DM, Mehrtash H. Development of measures for assessing mistreatment of women during facility-based childbirth based on labor observations. *BMJ Global Health* 2021;5:e004080.
28 WHO. Prevention and elimination of disrespect and abuse during childbirth [Internet]. WHO. World Health Organization, 2014. Available: http://www.who.int/reproductivehealth/topics/maternal_perinatal/statement-childbirth/en/ [Accessed 16 Nov 2020].
29 Maung TM, Mon NO, Mehrtash H, *et al*. Women's experiences of mistreatment during childbirth and their satisfaction with care: findings from a multicountry community-based study in four countries. *BMJ Glob Health* 2021;5:e003688.
30 Kujawski S, Mbaruku G, Freedman LP, *et al*. Association between Disrespect and abuse during childbirth and women's confidence in health facilities in Tanzania. *Matern Child Health J* 2015;19:2243–50.
31 Wilson M. *Constructing measures: an item response modeling approach*. New York, NY: Taylor & Francis Group, 2005.
32 Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care* 2000;38:II28–42.
33 Kleinman M, Teresi JA. Differential item functioning magnitude and impact measures from item response theory models. *Psychol Test Assess Model* 2016;58:79–98.
34 Masters GN, Wright BD. The Partial Credit Model. In: van der Linden WJ, Hambleton RK, eds. *Handbook of modern item response theory*. Springer New York, 1997: 101–21. http://link.springer.com/chapter/
35 Embretson SE. The new rules of measurement. *Psychol Assess* 1996;8:341–9.
36 Cappelleri JC, Jason Lundy J, Hays RD. Overview of classical test theory and item response theory for the quantitative assessment of

items in developing patient-reported outcomes measures. *Clin Ther* 2014;36:648–62.
37 Luciano JV, Ayuso-Mateos JL, Aguado J, *et al*. The 12-Item World Health organization disability assessment schedule II (WHO-DAS II): a nonparametric item response analysis. *BMC Med Res Methodol* 2010;10:45.
38 Prieto L, Alonso J, Lamarca R. Classical test theory versus Rasch analysis for quality of life questionnaire reduction. *Health Qual Life Outcomes* 2003;1:27.
39 Rehm J, Üstün TB, Saxena S, *et al*. On the development and psychometric testing of the who screening instrument to assess disablement in the general population. *Int J Methods Psychiatr Res* 1999;8:110–22.
40 Kim S, Moses T, Yoo HH. Effectiveness of item response theory (irt) proficiency estimation methods under adaptive multistage testing. *ETS Research Report Series* 2015;2015:1–19.
41 Strobl C, Kopf J, Zeileis A. Rasch trees: a new method for detecting differential item functioning in the Rasch model. *Psychometrika* 2015;80:289–316.
42 Kiefer T, Robitzsch A, Wu M. Test analysis modules [Internet], 2015. Available: cran.r-project.org/web/packages/TAM/TAM.pdf
43 Vedam S, Stoll K, Martin K, *et al*. The mother's autonomy in decision making (MADM) scale: Patient-led development and psychometric testing of a new instrument to evaluate experience of maternity care. *PLoS One* 2017;12:e0171804.
44 Kruk ME, Kujawski S, Mbaruku G, *et al*. Disrespectful and abusive treatment during facility delivery in Tanzania: a facility and community survey. *Health Policy Plan* 2018;33:e26–33.
45 Ford-Gilboe M, Wathen CN, Varcoe C, *et al*. Development of a brief measure of intimate partner violence experiences: the Composite Abuse Scale (Revised)-Short Form (CASR-SF). *BMJ Open* 2016;6:e012824.
46 Sharma G. An investigation into quality of care at the time of birth at public and private sector maternity facilities in Uttar Pradesh, India [Internet] [doctoral]. London School of Hygiene & Tropical Medicine, 2017. Available: https://researchonline.lshtm.ac.uk/id/eprint/4646087/ [Accessed 20 Feb 2020].
47 Warren CE, Njue R, Ndwiga C, *et al*. Manifestations and drivers of mistreatment of women during childbirth in Kenya: implications for measurement and developing interventions. *BMC Pregnancy Childbirth* 2017;17:102.
48 Cohen J, Rothschild C, Golub G, *et al*. Measuring the impact of cash transfers and behavioral 'nudges' on maternity care in Nairobi, Kenya. *Health Aff* 2017. ;;36:1956–64
49 Williams B, Coyle J, Healy D. The meaning of patient satisfaction: an explanation of high reported levels. *Soc Sci Med* 1998;47:1351–9.
50 Freedman LP, Kujawski SA, Mbuyita S, *et al*. Eye of the beholder? observation versus self-report in the measurement of disrespect and abuse during facility-based childbirth. *Reprod Health Matters* 2018;26:107–22.
51 Dey A, Shakya HB, Chandurkar D, *et al*. Discordance in self-report and observation data on mistreatment of women by providers during childbirth in Uttar Pradesh, India. *Reprod Health* 2017;14:149.
52 Roder-DeWan S, Gage AD, Hirschhorn LR, *et al*. Expectations of healthcare quality: a cross-sectional study of Internet users in 12 low- and middle-income countries. *PLoS Med* 2019;16:e1002879.
53 Diamond-Smith N, Treleaven E, Murthy N, *et al*. Women's empowerment and experiences of mistreatment during childbirth in facilities in Lucknow, India: results from a cross-sectional study. *BMC Pregnancy Childbirth* 2017;17:335.
54 Balde MD, Nasiri K, Mehrtash H, *et al*. Labour companionship and women's experiences of mistreatment during childbirth: results from a multi-country community-based survey. *BMJ Glob Health* 2020;5:e003564.