GLOBAL HEALTH ACTION, 2018
VOL. 11, 1445466
https://doi.org/10.1080/16549716.2018.1445466

ORIGINAL ARTICLE

OPEN ACCESS

Development and validation of the ExPRESS instrument for primary health care providers' evaluation of external supervision

Michael Schriver[a], Vincent Kalumire Cubaka[a,b], Peter Vedsted[c], Innocent Besigye[d] and Per Kallestrup[a]

[a]Center for Global Health, Department of Public Health, Aarhus University, Aarhus, Denmark; [b]School of Medicine and Pharmacy, College of Medicine and Health Sciences, University of Rwanda, Kigali, Rwanda; [c]Research Unit for General Practice, Department of Public Health, Aarhus University, Aarhus, Denmark; [d]Department of Family Medicine, School of Medicine, Makerere University, Kampala, Uganda

ABSTRACT

Background: External supervision of primary health care facilities to monitor and improve services is common in low-income countries. Currently there are no tools to measure the quality of support in external supervision in these countries.

Aim: To develop a provider-reported instrument to assess the support delivered through external supervision in Rwanda and other countries.

Methods: "External supervision: Provider Evaluation of Supervisor Support" (ExPRESS) was developed in 18 steps, primarily in Rwanda. Content validity was optimised using systematic search for related instruments, interviews, translations, and relevance assessments by international supervision experts as well as local experts in Nigeria, Kenya, Uganda and Rwanda. Construct validity and reliability were examined in two separate field tests, the first using exploratory factor analysis and a test–retest design, the second for confirmatory factor analysis.

Results: We included 16 items in section A ('*The most recent experience with an external supervisor*'), and 13 items in section B ('*The overall experience with external supervisors*'). Item-content validity index was acceptable. In field test I, test–retest had acceptable kappa values and exploratory factor analysis suggested relevant factors in sections A and B used for model hypotheses. In field test II, models were tested by confirmatory factor analysis fitting a 4-factor model for section A, and a 3-factor model for section B.

Conclusions: ExPRESS is a promising tool for evaluation of the quality of support of primary health care providers in external supervision of primary health care facilities in resource-constrained settings. ExPRESS may be used as specific feedback to external supervisors to help identify and address gaps in the supervision they provide. Further studies should determine optimal interpretation of scores and the number of respondents needed per supervisor to obtain precise results, as well as test the functionality of section B.

ARTICLE HISTORY
Received 24 October 2017
Accepted 21 February 2018

RESPONSIBLE EDITOR
Jennifer Stewart Williams, Umeå University, Sweden

KEYWORDS
Supportive supervision; instrument development; primary health care; Africa; Rwanda

Background

Health professionals in resource-constrained primary health care settings are likely to work in overburdened conditions, carry responsibilities above their level of training and receive little or no further clinical training or support [1–4]. Generally, supervision is regarded a core element to ensure high quality care [5]. The more remote the setting in which health professionals work, the higher the level of supervision needed [1].

In low-income countries, *external supervision* (i.e. supervision delivered by supervisors from outside the facility) of primary health care facilities appears to be common practice [6–9]. External supervision often focuses on management and administration more than on problem solving and feedback [6,7]. Yet, health policies across Africa describe support for providers' professional development as a component of external supervision [7,10–12], sometimes referred to as supportive supervision [8,13]. External supervisors may thus have a dual role that relates to: (1) managerial quality control of performance; and (2) formative support of providers. It has been suggested that there is a gap between health supervision policies and implementation of formative aspects of external supervision [7,14].

The external supportive supervision model [6–9,13] is described as unique to developing countries [15]. Numerous instruments have been developed in high-income settings to evaluate the quality of provider-centred supervision [16,17] and training [18] practices. The applicability of these instruments in management-centred, external supervision contexts has not been unexplored.

CONTACT Michael Schriver micschriver@gmail.com Råhøjvænget 13, 8260 Viby J, Denmark
Supplementary material for this article can be accessed here.

Questionnaire-based outcome measures applied in studies of external, supportive supervision in Africa are commonly non-validated [8].

Supervision context in Rwanda

In Rwanda, external supervisors regularly visit primary health care facilities (health centres) for evaluative and formative supervisory purposes [14,19]. The external supervisors work in teams under the district hospital to which health centres refer. Supervisors are typically clinically experienced nurses with a higher nursing degree [19]. One of the major supervision drivers is the monthly or quarterly performance evaluations, which constitute the core of a nationwide performance-based financing system [14].

The health centres have no medical doctors, and more than 90% of their providers are nurses with a basic secondary school-based nursing degree (known as an A2 degree). The providers do not have a personal supervisor. Supervision encounters may happen between one or more supervisors and one or more providers. The lack of a personal supervisor together with a high turnover, absenteeism and frequent provider shifts between services, make it likely that providers interact with a new supervisor at each supervision encounter [14,19].

A rating scale to assess external supervision may help assure supervision quality in these diverse contexts. Such an instrument should assess the construct *'Perceived quality of supportive aspects within external supervision of primary health care providers'*. It reflects a view of the provider as a direct beneficiary of external supervision despite its managerial and evaluative purposes [6,7].

Our aim was to develop a tool measuring provider-reported quality of supervision to be used to give feedback to supervisors and supervision teams in Rwanda to facilitate informed changes in the practice of external supervision [20]. Moreover, to empower providers with an opportunity to give feedback to supervisors within an otherwise asymmetric power relation [19]. The tool should thus focus on aspects of supervision potentially modifiable by supervisors, and cover key concepts in supportive supervision within health care. We aimed to make the tool applicable in other African countries.

Methods

Multiple methods were used. Table 1 gives an overview of 18 chronological steps in three phases in the development of the External supervision: Provider evaluation of supervisor support (ExPRESS) tool. While phase 0 and phase 1 represent a pre-designed logical order of steps, phase 2 represents additional steps that emerged as necessary or logical to address problems or shortages discovered during the development process. A detailed view of added, revised and removed items during these steps is included as supplementary material 1.

In this paper, item numbers corresponding to the questionnaire used in field test I (step 7) are referred to by small letters (a1–a16, b1–b13), and the item numbers in field test II (step 18) are referred to by capital letters (A1–A18, B1–B15).

Phase 0

In the preparatory phase, we conducted qualitative studies (step 1) to understand the practice of external supervision in Rwanda. We used focus group discussions with separate groups of providers and supervisors to explore the relationships between evaluative and formative supervision activities and between supervisors and providers. Methods and results are reported elsewhere [14,19].

We also conducted a systematic search (step 2) for published instruments measuring supervision or mentorship in health care to develop a bank of constructs and items (supplementary material 2 for search strategy). Further, we used reviews of directly or indirectly related instruments [16–18] and Google searches for non-published instruments. Additionally, we searched guidelines about supervision and mentoring within health or social sciences, and performed snowball searches in reference lists.

Phases 1 and 2

Conceptual model

The questionnaire is based on a reflective conceptual framework [21]. In the initial conceptual model (step 3) we categorised items according to Proctor's tripartition of supervisory tasks into normative (administration and performance evaluation), formative (education) and restorative (personal wellbeing at work) [22]. Further, we divided the questionnaire into a specific A and a generic B section as providers may interact with different supervisors from encounter to encounter. Section A evaluates the most recent supervision experience using items that providers may reasonably assess after each supervision encounter with an individual external supervisor. Section B represents a sum experience with external supervisors to ensure coverage.

Table 1. Phases and steps in the development of ExPRESS.

	Step	Objective	Description
Phase 0	1. Qualitative studies	Increase understanding of external supervision	Seven focus group discussions with providers and supervisors to understand experiences. Reported elsewhere.
	2. Instrument search	Identify supervision measurement instruments	Systematic search for instruments to measure supervision. No existing tool found applicable to external supervision context.
Phase 1	3. Conceptualisation I	Develop model for questionnaire	Defining construct. Categorisation in normative, formative and restorative functions. Division in section A (individual supervisor) and section B (supervision overall)
	4. Item development I	Develop item pool and adapt relevant items	Of > 400 items, 122 retained in item pool, of which six used directly, 22 modified or inspirational and eight new items added for the first version.
	5. Translation I	Prepare for tests in Rwanda	Forward and backward translation into Kinyarwanda
	6. Interviewing I	Cognitive testing of items	Individual cognitive interviews of 10 providers and one information expert. Two group discussion with five providers and six supervisors. 17 items modified, 10 items removed, 3 items added.
	7. Field test I	Factor structure and reliability	134 respondents, 58 retest. Exploratory factor analysis and test–retest reliability
Phase 2	8. Conceptualisation II	Refine conceptual model	Systematic refinement of conceptualisation using multiple sources on supportive supervision
	9. Item development II	Adapt to refined model	11 items modified, 1 item removed, 5 items added
	10. Interviewing II	Lexical test	Two interviews with professional native English linguist to test lexical qualities of English version
	11. Relevance assessment I	Content validation by international expert	Relevance assessment of items by four international experts on supportive supervision analysed via the Content Validity Index. 14 items modified, 5 items removed, 6 items added
	12. Item development III	Review and revise prior to new translation	12 items modified, 4 items added.
	13. Translation II	Prepare final Rwandan version	Renewed translation and back-translation of all items due to several changes and modifications
	14. Relevance assessment II	Content validation by regional experts	Relevance assessment of items by five providers and five external supervisors in Rwanda, Uganda, Kenya and Nigeria analysed via the Content Validity Index
	15. Interviewing III	Testing response scale	Two group discussions with five providers on the response scale. Response scale changed for section B.
	16. Item development IV	Adding latent variable	Adding three items of a latent variable 'Solving problems jointly'
	17. Translation III	Translate added items	Translation and back-translation of added items for the 'Joint Problem Solving' latent variable.
	18. Field test II	Confirmatory factor analysis	154 respondents. Confirmatory factor analysis and Differential Item Functioning

In phase 2, we refined the conceptual model (step 8) using key articles and guidelines on supportive supervision in a low-resource setting [13,23–30]. Supportive supervision contents were extracted, discussed and categorised, leading to a list of key aspects to cover in the questionnaire (supplementary material 3).

Item development

Two researchers (MS and VKC, in step 4) screened all items identified in the literature search and created an item pool of those appropriate in contexts where:

- Providers may not have a personal supervisor
- The supervisor is from an external institution
- Supervisors may carry a managerial role

Further, each item should:

- Focus on a specific event related to supervision
- Use simple, non-idiomatic phrases

Items in the pool were inductively categorised in themes. Relevance to the instrument construct was assessed as ”yes”, ”no” and ”maybe” by two researchers (MS and VKC) independently. Subsequently, an iterative process of discussion among researchers informed by qualitative findings [14,19], supervision literature, conceptual models, item categories and considerations of language, semantics and level of specification, led to the composition of a first combination of items for section A and B to undergo a translation. Following the refined conceptual model, the combination of items was again modified (See Table 1 step 9, and supplementary material 1).

Items were developed with focus on both clinical and non-clinical aspects, as both may be supervised in the same encounter. It was difficult to find appropriate items to evaluate the key concept of joint problem solving. At an advanced stage (step 16), a publication [31] provided an idea for how to add ‘solving problems jointly’ as a latent variable (a variable that may not be directly observed but may be indirectly measured through a set of observable items) in section A, using phrases such as ‘engaged me in’ and ‘involved me in’.

Translation

Items were developed in English and translated into Kinyarwanda for testing in Rwanda. We followed a standardized approach [32]. Two translators, a professional translator not knowledgeable about supervision and someone who had published articles about health care supervision in Rwanda, did the translation of items into Kinyarwanda. Two other translators, a native English speaker and someone who spoke English as a second language from early childhood, did the back translation. To obtain consensus of the translation of each item, MS and VKC met with the first translators, and subsequently with all four translators. As items were translated during the development of the instrument, complete translation and back-translation including meetings was done twice (steps 5 and 13). Subsequent addition of a latent variable (‘solving problems jointly’) required a third translation process of three items (step 17), with participation of only one back-translator.

When discussions suggested a need to change the original English version, this was done only if there was consensus between the two researchers and the translators.

Interviews

For cognitive testing of items (step 6) we used a combination of ”Think aloud” and ”Probing” techniques [33] (supplementary material 4 for interview guide). Initially, a local communication expert and a local external supervisor were interviewed, followed by 10 individual interviews with local primary health care providers at health centres. Interviews were held in Kinyarwanda by a trained interviewer with a social science background, who also took notes, item by item. Interviews lasted 1.5–2 hours and were not recorded. After each interview, notes were discussed between the interviewer, MS and VKC, and agreed changes were applied to ExPRESS before the next cognitive test interview. Further, two focus group discussions facilitated by the same interviewer, one with six providers (five females, one male) and one with five external supervisors (three males, two females), examined meaning and relevance item by item, and suggested missing concepts. Interviews and focus group discussions led to several changes of items (see supplementary material 1).

Response scale

Initially, we used a 5-point neutral-centred agreement response scale with the advantage of uniform applicability regardless of whether items are phrased positively or negatively. In four initial, cognitive interviews (step 6) the most positive response option was endorsed for nearly all items, and interviewees did not endorse negative response options. This was in spite of the providers verbally criticising their supervisors on the same items. Therefore, a 5-point quality response scale was applied instead: ‘1 = poor; 2 = fair; 3 = good; 4 = very good; 5 = excellent’, to expand the positive spectrum. Interviewees did not report problems with understanding or using this scale.

In step 15, we held two focus group discussions each with five primary health care providers to discuss alternative response scales. For section A, the quality response scale described above, a variant of the quality scale and a 4-point scale ('no, not at all', 'yes, a little', 'yes, somewhat', 'yes, very much') were explored. For section B, we explored the quality response scale and a frequency scale ('never', 'sometimes', 'usually', 'quite often', 'always'). First, providers individually chose their preferred scale, and then discussed their preferences. All preferred the quality response scale (poor–excellent) for section A. Due to time-related items in section B, most but not all preferred the frequency response scale, which was applied in field test II.

Data collection in field tests I and II

Questionnaires in field tests I and II were self-administered after brief, face-to-face information by one of two trained assistants. For factor analysis we needed four respondents per item and for test–retest 50 respondents, as recommended [34]. We added 15–20% more respondents due to anticipated missing items. In field test I, all respondents were nurses recruited at their health centre after agreement with facility managers. In field test II, 107 (69%) nurses were recruited in this way, and the rest were nurses recruited from nursing schools, where they attended further training while being employed at a health centre. Only respondents who had experienced external supervision in the previous four months were invited. Participants filled in the questionnaire in privacy. All data in field tests I and II were entered into EpiData 2.0.5.17 using double entry, and analysed in STATA 14.2.

Field test I

The purpose of field test I (step 7) was to explore structural validity (the combination of items that would adequately reflect the construct of the questionnaire), and to conduct a test–retest reliability study (testing to what extent a provider would give the same responses about the same supervision experience when asked at two different moments in time). Structural validity was assessed with explorative factor analysis (EFA), in which factor loadings are used to study the correlation of items. The purpose of this is to identify a meaningful categorisation of items in which each item has a high factor loading with only one group of items, and thus does not cross-load (correlate) with other groups of items. We used so-called polychoric correlation matrices [34], principal axis factoring and promax oblique rotation [35]. We considered a factor loading ±0.50 or higher as practically significant, and only explored loadings ±0.30 or higher [36]. Loadings and cross-loadings ±0.30 to ±0.49 were considered potentially problematic.

First, a forced 2-factor structure for the entire questionnaire (sections A and B) was explored. Secondly, structural validity was assessed within section A and B, respectively. Here, number of factors were explored stepwise, starting with the maximum potential factors as suggested by scree plot, the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and eigenvalues, until at least two and preferably three items loaded 0.5 or greater on all factors [35,37,38].

For test–retest reliability, we considered respondents 'stable' if they had not experienced supervision between the first and second time they filled in the questionnaire. The time between responses was 12–14 days. We used weighted Cohen's kappa [39] with linear and quadratic weights [40]. Additionally, a modified weight of identical answers as 1, directly adjacent as 0.8 and all others as 0 was used, since we expected the majority of retest responses to be within ± 1 of the test response. We applied Landis and Koch for kappa-values: 0–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1 nearly perfect [34].

Content validation

We conducted relevance assessments (steps 11 and 14) using the Content Validity Index (CVI) to get a consensus estimate. Each item was scored by experts as: (1) *not at all*; (2) *somewhat*; (3) *quite*; or (4) *highly relevant* for measuring a given construct. The item-CVI is the fraction of experts who found an item highly or quite relevant. With five or fewer experts, the item-CVI should be 1 (relevant to all experts) to retain an item, whereas for more than five experts an item-CVI above 0.78 was considered acceptable [41,42].

Four international experts on supportive supervision in Sub-Saharan Africa who had published on the matter [13,27,43,44], assessed the relevance of items. We knew of these experts only through their publications, which we considered important points of reference for supportive supervision in Africa. All were contacted by email. For items considered *somewhat* or *not at all relevant*, experts explained their score to enable discovery of solutions to item problems. After modifications, experts re-assessed the items [42].

Subsequently (step 14), we conducted a content validity assessment of the revised English questionnaire in Nigeria, Uganda, Kenya and Rwanda, by five external supervisors and five primary health care providers in each country (40 individuals). A collaborator in each country helped to collect the data using a standardized relevance assessment questionnaire. Respondents had to be able to read, write and understand English. Further, they were required to have a minimum of two years of experience as a provider in or a supervisor of public non-hospital primary health care facilities, as well as have visited a primary health care facility to supervise (supervisors), or experienced external supervision at their facility (providers) within the previous four months.

Field test II

In field test II (step 18) we conducted a CFA for section A and B separately, using maximum likelihood with the Satorra-Bentler (SB) estimation, which is robust to non-normality [45,46]. This was relevant since neither section showed multivariate normality. Model fit was considered good with a p-value for the chi^2 test > 0.01, Tucker Lewis Index (TLI) and Comparative Fit Index (CFI) > 0.95, root mean square error of approximation (RMSEA) < 0.6 and standardized root mean square residual (SRMR) < 0.8, and acceptable if approaching these values [34,47–49]. Since SRMR and the confidence interval of RMSEA were not available for the SB estimation, these are reported based on full (non-SB) maximum likelihood.

For section A, we hypothesised a 4-factor structure as the best fit. This was informed by a 3-factor output of the EFA and a fourth latent variable *'solving problems jointly'* was added later. To compare the fit, we had predefined relevant 3- and 2-factor models. As indicated by results of the EFA, we hypothesised that the fit could potentially be improved by removing A2 and moving A9 to '*Generating comfort*' (see Table 5).

For section B we hypothesised that a 4-factor structure would be the best fit, and had predefined relevant 2-factor, 3-factor, 4-factor models to compare fit, as well as improving fit by excluding items B3, B5 and B15 (see Table 5). The hypothesised models are included as supplementary material 5.

Results

Phase 0

The systematic search identified 21 measurement instruments related to supervision of which five were not published in scientific journals [50–54], six were published but not as papers to validate the instrument [55–60] and 11 were published as validation studies [61–71]. Additionally, three instruments were found where respondents assessed an external event [18,72,73], and one instrument related to the primary health care field [74]. Over 400 items were identified, and 122 were retained in the item pool. The most common reasons for excluding items were that they were complicated in phrasing, vague, or inappropriate for the context of external supervision.

Phase 1

Item development

Following categorisation of pooled items as well as discussion and assessment of their relevance, a first version of the questionnaire was composed for cognitive testing. For section A, four items from the item pool were used with no modifications, 14 items were modified or the idea was used to develop another item, and based on qualitative supervision data, three new items were added [14,19]. For section B, two items were used without modifications, eight items were modified or the idea was used to develop a new item, and five new items were added. After cognitive testing, 10 items were removed, three items were added and 17 items modified. After the refined conceptual model, one item was removed, five items added and 11 items were modified (supplementary material 1).

Table 2. Data quality and characteristics presented as range and (mean) across all items within a section.

	Field test I (N = 134)		Field test II (N = 158)	
Parameters	Section A	Section B	Section A	Section B
Item mean	2.6–4.0 (3.5)	2.8–4.0 (3.5)	2.6–4 (3.3)	2.8–4.3 (3.7)
Item SD	0.9–1.2 (1.1)	0.8–1.2 (1.0)	0.9–1.2 (1.1)	0.9–1.4 (1.1)
Item median	3–4 (3.7)	3–4 (3.7)	3–4 (3.4)	2–5 (3.8)
% with lowest response in item	1–25 (6)	1–12 (4)	2–25 (10)	1–29 (5.8)
% with highest response in item	6–33 (17)	6–31 (19)	3–29 (11)	9–60 (35)
Item kurtosis	1.8–3.6 (2.7)	2.0–3.3 (2.5)	1.9–4.0 (2.7)	1.6–4.0 (2.7)
Item skewness, absolute values	0.2–0.9 (0.5)	0.1–0.7 (0.3)	0.0–0.9 (0.5)	0.2–1.3 (0.8)

OBS: Items of field test I and II are different

Field test I

All invited participants responded to the field test I questionnaire version (items a1–a16 and b1–b13, supplementary material 1). A total of 134 primary health care nurses, 52% of whom were from districts in the capital, Kigali, participated. Respondents were from 27 health centres, 52% had their most recent supervision within the previous month, and 75% were female, reflecting a predominance of females in the nursing profession (supplementary material 6 for participant characteristics). Respondents had assessed 36 different supervisors in section A (24 did not provide the supervisor name). A total of 111 and 119 respondents had no missing items for section A and section B, respectively. Table 2 shows the range and mean of various descriptive statistical indicators across all items within a section (including field test II).

Exploratory factor analysis. In the forced 2-factor structure, all section A items loaded above 0.50 in factor 1 (except a13 loading 0.47) and all section B items loaded above 0.50 in factor 2. Only one item (b13) cross-loaded above 0.3 (0.34).

For section A, up to six factors were suggested. Following stepwise exploration of loadings, we found a potential fit of a 3-factor model corresponding to ‘*Generating comfort*’, ‘*Understanding work of providers*’ and ‘*Building provider capacity*’, retaining items a1–a12. Item a1 had lowest loading (0.56) and communality (0.53). Item a7 cross-loaded with factors of ‘*Generating comfort*’ and ‘*Understanding work*’ in several models. These observations of a1 (=A2 in field test II) and a7 (= A9 in field test II) were considered for the CFA models in field test II. Item a13 had loadings and communality below 0.5. Due to content validity it was moved to section B instead of being excluded. Items a14–a16 were excluded as they did not represent specific supervisory events, and loaded on a fourth factor with which several items cross-loaded.

For section B, up to seven factors were suggested. Using stepwise exploration, a 4-factor model emerged with factors corresponding to ‘*Planning*’, ‘*Team work*’, ‘*Assessing Performance*’ and ‘*Capacity to teach*’, retaining items b1-b11. Items b13 (≈ B15 in field test II) and b12 loaded on a factor with several cross-loadings, and did not evaluate specific supervisory events. Items b3 (≈ B5 in field test II) and b10 (= B3 in field test II) loaded below 0.5. These were considered for CFA modelling in field test II.

Table 3. Test–retest differences and kappa values of field test I (N = 58).

	Differences (Retest minus test)										Weighted Kappa		
Item	−4	−3	−2	−1	0	1	2	3	4	Missing	Modified	Linear	Quadratic
a1			1	12	31	12		1		1	0,63	0,47	0,59
a2		1	3	15	27	8	1			3	0,57	0,47	0,60
a3			4	11	32	8	1			2	0,63	0,58	0,71
a4			4	17	27	7	2			1	0,64	0,60	0,72
a5			1	8	40	7	1			1	0,75	0,68	0,78
a6			4	15	26	10	3			0	0,45	0,39	0,55
a7		1	1	13	29	11	1			2	0,61	0,48	0,63
a8			5	10	31	11	1			0	0,59	0,53	0,67
a9		1	3	17	24	8	3	2		0	0,47	0,39	0,52
a10			3	12	30	7	4			2	0,50	0,46	0,57
a11		1	1	13	28	7	1	2		5	0,61	0,50	0,59
a12	2		1	13	32	8	1			1	0,66	0,53	0,57
a13	1	1	4	8	31	11		1		1	0,61	0,53	0,60
a14			3	12	36	6	1			0	0,65	0,58	0,69
a15			3	20	28	7				0	0,59	0,45	0,62
a16		1	2	17	30	6	1			1	0,60	0,50	0,63
% of total	0%	1%	5%	23%	53%	15%	2%	1%	0%				
b1			4	10	32	6	4		1	1	0,49	0,47	0,51
b2			4	13	23	12	5		1	0	0,39	0,33	0,45
b3			1	12	27	13	2	1		2	0,55	0,42	0,55
b4		1	4	11	35	5	1			1	0,59	0,55	0,61
b5			6	12	27	10	2			1	0,42	0,38	0,50
b6		1	5	10	29	10	3			0	0,48	0,44	0,55
b7			5	12	32	6	3			0	0,47	0,46	0,55
b8		1	3	13	30	9	1			1	0,56	0,47	0,59
b9			2	17	29	6	3			1	0,54	0,46	0,60
b10		1		11	32	11	2			1	0,55	0,43	0,50
b11			4	17	27	5	3	1		1	0,42	0,37	0,48
b12			2	17	35	3				1	0,72	0,61	0,74
b13			4	15	33	6				0	0,57	0,49	0,59
% of total	0%	1%	6%	23%	53%	14%	4%	0%	0%				

Item numbers correspond to field test I version, non-corresponding with item numbers of field test II

Test–retest reliability. Of 134 providers in field test I, 58 had not experienced supervision since their response and participated in the retest (supplementary material 6 for characteristics). Table 3 shows the distribution of differences in test and retest responses, number of missing responses per item and weighted kappa values.

More than 90% of all retest responses were within +/- 1 of the test response. In all cases, linear weights had the lowest kappa values, and in most cases, quadratic weights the highest. With the suggested modified weight, all items had moderate to substantial agreement, except b2 with κ = 0.39.

Phase 2

Content validation by experts using the CVI

Following relevance assessment by four international experts in supportive supervision, we deleted five, modified 14 and added six items (see supplementary material 1). New and modified items were subsequently assessed by the same experts, as a 2nd iteration [42]. Here, only item B1 had an item-CVI below 1 (supplementary material 7). The item was included for field testing due to relevance in the qualitative studies.

The regional relevance assessment by five supervisors and five providers in each of four countries had acceptable item-CVI for all items except in Nigeria for item A2, A17, and a previous version of item A7 (Supplementary material 7). In Rwanda, these items were found relevant, and therefore included in field test II.

Field study II

Among 154 respondents, 72% were female, 90% had more than three years of practice experience and 68% had their most recent supervision within the previous two months (supplementary material 6 for participant characteristics). Respondents came from 17 different districts, and had evaluated 69 different supervisors in section A (eight respondents had not reported the supervisor name). Of 154 respondents, 146 were retained for CFA of section A and 145 for section B, as they had no missing items. Table 2 shows that 35% of respondents endorsed the highest possible response ('always') in section B items, compared to 11% ('excellent') in section A items. Item B1 (see Table 5) was included in the field test despite a CVI of 0.75 and had the lowest median and mean suggesting that it reflected a perceived problem. Table 4 shows goodness of fit output of the confirmatory factor analysis.

A reasonable fit was found for section A with the hypothesised 4-factor model, improved by excluding item A2 and moving A9 to factor 1 as hypothesised. The model improved by adding error correlations between items A3 and A4, and items A16 and A17, which was not predicted. Conceptually, these error correlations were reasonable and did not indicate redundancy.

Item A13 ('followed up on previous discussions') had a loading of 0.51 and was previously found irrelevant by an international expert (see supplementary material 1). It was therefore discussed and found inappropriate for section A, not necessarily linked to support and therefore excluded. Figure 1 shows the final 4-factor, 16-item model.

For section B, excluding item B3 significantly improved the fit of the proposed 4-factor model. Item B15 was non-specific and somewhat abstract, and was excluded to slightly improve fit. While

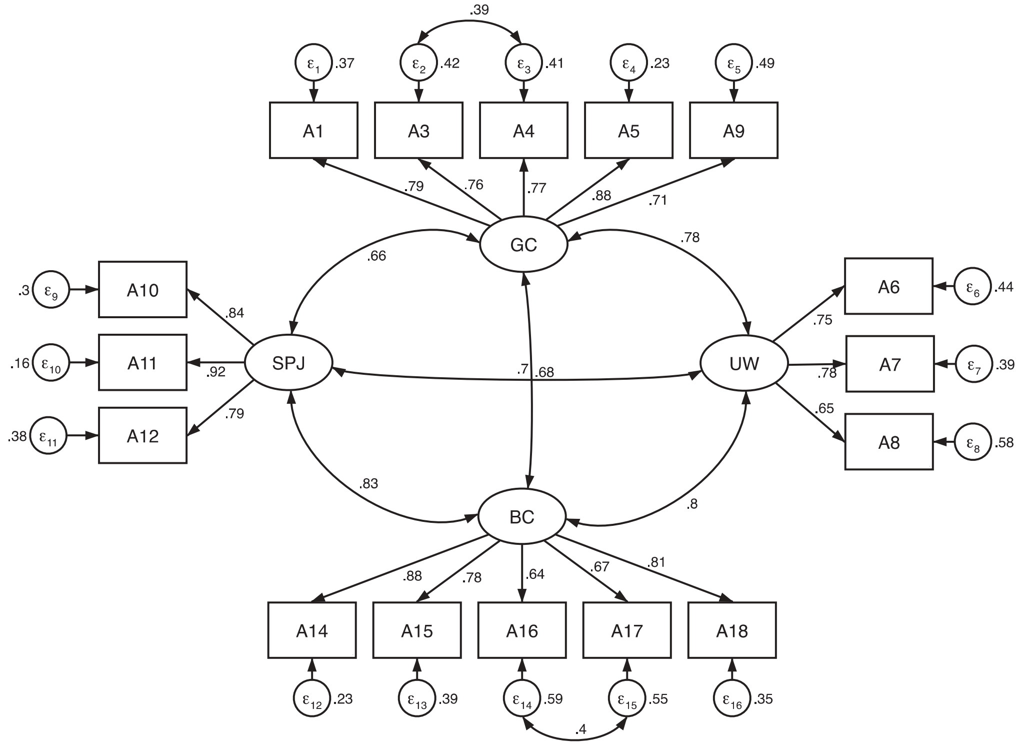
Table 4. Confirmatory factor analysis in field test II. Model fit of section A and B.

Model	Chi2*	df*	p*	CFI*	TLI*	RMSEA*	RMSEA (90% CI)		SRMR
A: 1 factor	375	153	0,00	0,81	0,78	0,11	0,13	(0,12; 0,15)	0,08
A: 2 factors	363	153	0,00	0,82	0,79	0,11	0,13	(0,12; 0,14)	0,08
A: 3 factors	266	153	0,00	0,89	0,88	0,08	0,10	(0,09; 0,12)	0,08
A: 4 factors	206	153	0,00	0,94	0,93	0,06	0,09	(0,07; 0,10)	0,06
A: 4 factors, exclude A2	177	136	0,00	0,95	0,94	0,06	0,08	(0,07; 0,10)	0,06
A: 4 factors, exclude A2, move A9	162	136	0,00	0,96	0,95	0,05	0,08	(0,06; 0,09)	0,05
A: 4 factors, exclude A2, A13, move A9	149	120	0,00	0,96	0,95	0,06	0,08	(0,06; 0,10)	0,05
A: As above + error correlations**	119	120	0,05	0,98	0,98	0,04	0,06	(0,04; 0,08)	0,04
B: 1 factor	182	105	0,00	0,86	0,83	0,08	0,10	(0,08; 0,12)	0,07
B: 2 factors	168	105	0,00	0,88	0,85	0,08	0,09	(0,08; 0,11)	0,07
B: 3 factors	150	105	0,00	0,90	0,88	0,07	0,09	(0,07; 0,11)	0,06
B: 4 factors	146	105	0,00	0,90	0,88	0,07	0,09	(0,07; 0,11)	0,06
B: 4 factors, exclude B3	107	91	0,00	0,94	0,92	0,06	0,07	(0,05; 0,09)	0,05
B: 4 factors, exclude B3 B15	86	78	0,01	0,95	0,93	0,06	0,07	(0,04; 0,09)	0,05
B: 4 factors, exclude B3 B15 B5	69	66	0,03	0,96	0,95	0,05	0,07	(0,04; 0,09)	0,05
B: 3 factors, exclude B3 B15 B5	76	66	0,01	0,95	0,94	0,06	0,07	(0,04; 0,09)	0,05
B: 3 factors, exclude B3 B15	75	66	0,02	0,95	0,94	0,06	0,07	(0,04; 0,09)	0,06

Df: Degrees of freedom, CFI: Confirmatory fit index, TLI: Tucker-Lewis Index, RMSEA: root mean square error of approximation, SRMR: standardized root mean square residual.

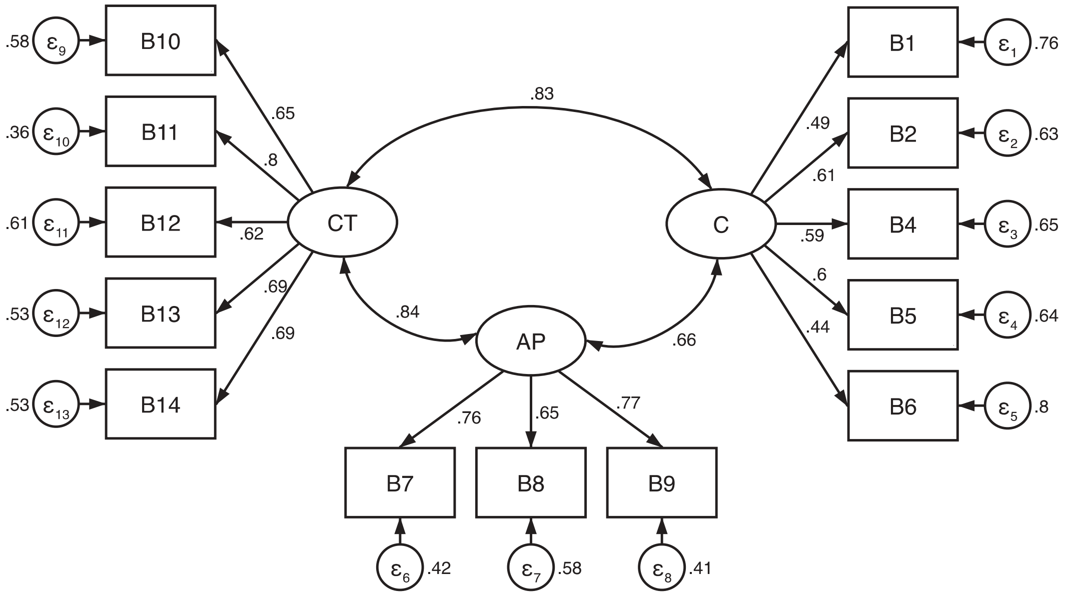
*Based on Satorra-Bentler estimation (not available for confidence intervals of RMSEA nor for SRMR).

** Error correlations: A3–A4 and A16–A17. Bold model: the final selected model.



GC: Generating comfort; UW: Understanding work, SPJ Solving problems jointly,
BC: Building capacity

Figure 1. Final structural equation model for section A with standardised factor loadings, error terms and error correlations.
GC: Generating comfort, UW: Understanding work, SPJ: Solving problems jointly, BC: Building capacity



C: Collaborating; AP: Assessing performance; CT: Capacity to teach

Figure 2. Final structural equation model for section B with standardised factor loadings and error terms.
C: Collaborating; AP: Assessing performance; CT: Capacity to teach.

excluding B5 slightly improved fit, it was retained for content validity reasons. To avoid a factor of two items we adopted the 3-factor comparison model, which also had an appropriate fit. Improvements from error correlations were not conceptually appropriate. The final 3-factor and 13-item model is shown in Figure 2.

Cronbach's alpha was 0.93 for the final 16-item version of section A with item-rest correlations of 0.55 to 0.74. The final 13-item version of section B

Table 5. Final items of ExPRESS. *Items removed following field test II are indicated with *.*

	Latent variable
Section A. During the most recent supervision, the supervisor…	
A1. Communicated in a friendly way	GC
*A2. Explained the purpose of the supervision visit	
A3. Wanted to know my opinions	GC
A4. Listened to me attentively	GC
A5. Treated me with respect	GC
A6. Observed how I carry out specific tasks of my work	UW
A7. Spend enough time discussing my work tasks with me	UW
A8. Was familiar with my area of work	UW
A9. Showed appreciation for my work	GC
A10. Asked me what problems I experience at work	SPJ
A11. Engaged me in discussions to examine problems at work	SPJ
A12. Involved me in deciding how to handle problems at work	SPJ
*A13. Followed up on previous discussions	
A14. Encouraged me to ask questions	BC
A15. Gave useful feedback about my work	BC
A16. Asked me what I need to learn more about	BC
A17. Discussed next steps	BC
A18. Checked to make sure I understood everything we discussed	BC
Section B. In general, supervisors…	
B1. Keep their supervision appointments	C
B2. Try not to disturb patient care	C
*B3. Treat women and men equally	
B4. Gather me and my colleagues for discussing as a group, when needed	C
B5. Maintain proper confidentiality of work-related information	C
B6. Strengthen the teamwork at my facility, when needed	C
B7. Explain the criteria used when assessing my performance	AP
B8. Assess my performance in a fair way	AP
B9. Give useful feedback after assessing my performance	AP
B10. Have sufficient clinical skills and knowledge	CT
B11. Explain difficult issues in a clear way	CT
B12. Update me when there are major changes in guidelines	CT
B13. Help make sure my needs for training are met	CT
B14. Help me feel confident at work	CT
*B15. Conduct supervision in a way that makes me provide better care	

The instrument has the following latent variables:
GC: Generating comfort, UW: Understanding work, SPJ: Solving problems jointly, BC: Building capacity, C: Collaborating, AP: Assessing performance, CT: Supervisor capacity to teach.

had alpha 0.87 and item-rest correlations from 0.39 to 0.70.

The final questionnaire is prese final questionnaire is prese final questionnaire is presented in Table 5.

The individual supervisor at a specific supervision encounter is assessed in section A, which contains the latent variables *generating comfort* (5 items), *understanding work* (3 items), *solving problems jointly* (3 items) and *building capacity* (5 items). The overall experience of supervision is assessed in section B, which contains the latent variables *collaborating* (5 items), *assessing performance* (3 items) and *capacity to teach* (5 items).

Discussion

This study documents the rigorous process of development and validation of the ExPRESS questionnaire using multiple strategies to allow for triangulation. Items were developed through an iterative approach using an item pool derived from 25 existing instruments, and discussions informed by the construct, conceptual framework and qualitative supervision data grounded in the experiences and perceptions of primary health care providers and their supervisors. A standardized translation process, cognitive interviewing and lexical testing resulted in several relevant modifications. Further modifications were made following content validation using the content validity index among international experts as well as among supervisors and primary health care providers in other sub-Saharan African countries. Structural validation was conducted using EFA in field test I, which guided further instrument development and generation of model hypotheses tested in field test II.

Contribution to supervision measurement

To our knowledge, ExPRESS is the only instrument designed and validated for primary health care providers to evaluate the quality of support in *external* supervision, in which normative functions such as performance control generally dominate. While the tools retrieved for this study assumed a provider-centred supervision approach (with some exceptions [52,60]), ExPRESS is appropriate for managerial supervision that claims to maintain provider support as a key objective. This form of supervision is particularly prevalent in resource-constrained settings.

Some existing tools evaluate a specific encounter [51,52,58,61,71], and others a sum of experiences [57,59,60,62–70], although this may not be explicit. ExPRESS is the only instrument divided into two sections to assess both a specific encounter and a sum experience of supervision. This is relevant for diversified external supervision contexts where providers may encounter different supervisors.

The items included in ExPRESS generally assess specific events that may or may not take place in the encounter between a provider and a supervisor. This event-orientation of items allows ExPRESS to provide concrete feedback to a named supervisor and/or a supervisory team on areas to improve. Only one tool [52] specified the particular supervision encounter assessed and used event-oriented items, but was neither developed for administration by supervisees nor validated.

Scoring and interpretation

Optimal scoring and interpretation of the instrument remain to be determined. Using scores 1 (lowest) through 5 (highest) as response options, we preliminarily suggest that scores below 80% of the maximum possible score (corresponding to the three lowest response options, if each item is considered separately) indicate a practical need for improvement. This threshold could also be used for items combined. For instance the latent variable of three items '*solving problems jointly*' would have a maximum possible score of 5*3 = 15, and thus a score of 11 or below would indicate a need for improvement. In case of missing items, the maximum possible score would be altered (by subtracting 5 per item missing) and the score needed for a proportion of a minimum of 80% would thus be proportionately altered [75].

Criterion validation could be possible using other measures of supervision and achieved competences, and construct validity may be further evaluated by 'known group' analysis and item response theory. Further studies are needed to determine the number of assessments necessary per supervisor in section A and per supervision team in section B for achieving appropriate statistical precision. Comparable instruments recommend from 4 [76] to 20 [72] assessments per evaluatee.

ExPRESS is a measure of providers' *expression* of supervisors' behaviour. It should not be interpreted as a measure of supervisor behaviour [77]. Perceptions of the same supervision event may differ between people depending on their personality [78].

Strengths and limitations

This study has a number of strengths. The design involved multiple phases and methods including systematic search, qualitative explorations and mirroring steps of item development, content validation and structural validation, leading to relevant modifications throughout the process. By developing the tool in English with the purpose of making it useful across contexts of external supervision and using a standardized translation process, we avoided local language issues and idioms while ensuring cultural and contextual adaptation. Regional relevance assessments indicated high generalizability, and international experts were involved to improve as well as assess the instrument. We also reached the intended number of respondents for the test–retest, field tests I and II, and respondents represented districts and health centres across Rwanda.

The study has several limitations related to the design, data collection and data analysis. ExPRESS was framed as a reflective measurement model with latent variables reflecting supervisor traits and abilities. However, this is not self-evident and the event-orientation of items could raise reasonable arguments for formative relationships [21]. The responsiveness of ExPRESS, that is its ability to measure change over time, was not evaluated, but would be needed to apply ExPRESS in measuring effects of supervision interventions.

Since the main part of the cognitive testing was conducted on preliminary versions of the questionnaire during phase 1, items A10-A12 did not undergo cognitive or relevance testing in their final form. However, as they did not have higher missing rates than other items in field test II and represented modifications of items previously tested and found relevant, we considered their content validity acceptable.

Test–retest reliability data was collected in field test I, which may not be transferred to the final questionnaire version. While field test I data was collected from providers during the daytime and at health centres, this was not feasible for the retest data two weeks later, which for many was collected in the evening or outside the health centre. This may have caused an underestimation of agreement between test and retest [34]. Finally, in field test II we collected data on a frequency response scale for section B, as opposed to field test I, where a quality response scale was used. This may in part explain the significant difference in the percentage endorsing the highest 5-point response. A further study may establish the extent to which the frequency response scale contributes to a ceiling effect compared to the quality response scale.

Applying a 5-point ordinal scale as continuous data in CFA and using the maximum likelihood method has been shown to be appropriate [49]. The risk is to wrongly reject a proper model (type 1 error) [45]. We used the SB estimation due to questionable normality of section B in particular. The asymptotic distribution free method is applicable for non-continuous data, but was not applied as it may reject properly specified models if sample sizes are small ($N < 500$) or deviation from normality is minimal [45]. It has been suggested that non-normality is not problematic for the maximum likelihood method until univariate skewness and kurtosis approach 2.0 and 10.0, respectively [45]; our data is below these limits (Table 2).

Recall bias may be a concern for the section A assessing the most recent supervision. Therefore, we tried to identify participants who were recently supervised. In field test I, almost 50% and in field test II almost 80% of respondents had their most recent supervision experience over a month before answering the questionnaire. Therefore, the assessment may be hampered by recall bias. On the other hand, a more precise measure of an experience may require time to consider the experience [79,80]. We found measurement invariance for all items when comparing respondents supervised more and less than one month prior to the field test.

Conclusion

External supervision is a common strategy in primary health care management in resource-constrained settings. This paper presents the stepwise development of a novel instrument, ExPRESS, to measure the quality of support delivered through external supervision as assessed by its direct beneficiaries – primary health care providers. The instrument includes a section A assessing an individual external supervisor at a specific supervisory encounter, and a section B assessing external supervisors in general. Items were found relevant by experts of supportive supervision, as well as by providers and supervisors in four African countries.

We believe ExPRESS has a high content validity and a reasonable structural validity, and can be useful to evaluate external supervision in resource-constrained primary health care settings. This may include under-resourced settings in high-income countries. It is freely available to collaborators for non-commercial use. Further analyses must focus on scoring, interpretation, responsiveness and using the tool for feedback as well as on setting up a database of representative samples to explore how ExPRESS evaluates the quality of external supervision.

Acknowledgments

We kindly thank our sponsors the Karen Elise Jensen Foundation. We thank participants and collaborators at all levels of the ExPRESS development process for their help in shaping and improving the instrument. Specifically, we owe gratitude to several individuals, each an expert in her/his own right, who assisted us in organisational or scientific matters. Their names in alphabetic order are: Bruno Uzabakiriho, Calvin Wilson, Danjuma Ayotunde Bello, Ditte Lystbaek Andreasen, Eva Ørnbøl, Eric Sarriot, Jean Babtiste Kakoma, Jeanvier Kayitare Bananeza, John Doldo IV, Jørn Atterman, Judith Karia, Lars Bønløkke, Lani Marquez, Maaike Flinkenflögel, Manzi Anatole, Mieke Visser, Michael Miller, Patrick Kyamanywa, Phillip Cotton, Qyle Jro, Rakshika Bhana, Sylvere Itangishaka and Uchenna Anozie.

Author contributions

Idea: MS, VKC. Study design: MS, VKC, PV, PK. Collection of Data: MS, VKC, IB. Analysis of data: MS, VKC, IB, PV, PK. First manuscript draft: MS. Manuscript review: MS, VKC, IB, PV, PK.

Disclosure statement

No potential conflict of interest was reported by the authors.

Ethics and consent

The studies were approved by the Institutional Review Board of the University of Rwanda (letters: 13/2014; 216/2015 and 324/2016). Participants in field tests, interviews and focus group discussions provided written informed consent.

Funding information

This project was funded by the Karen Elise Jensen Foundation in Denmark.

Paper context

Worldwide, primary health care providers experience visits at their health facilities from external supervisors with the purpose to improve quality of care. However, no tool has been validated to assess the quality of the support provided in such external supervision visits. This study offers a way to bridge this gap by developing a provider-reported tool to serve as specific feedback to external supervisors, developed for applicability in resource-constrained primary health care settings worldwide, and validated for use in Rwanda.

ORCID

Michael Schriver http://orcid.org/0000-0002-8778-2599
Vincent Kalumire Cubaka http://orcid.org/0000-0001-7449-2421
Peter Vedsted http://orcid.org/0000-0003-2113-5599
Per Kallestrup http://orcid.org/0000-0001-6041-4510

References

[1] All-Party Parliamentary Group on Global Health, Africa All-Party Parliamentary Group. All the talents [Internet]. 2012. Available from: www.appg-globalhealth.org.uk

[2] Moosa S, Wojczewski S, Hoffmann K, et al. The inverse primary care law in sub-Saharan Africa. Britiish J Gen Pract. 2014 June;64(623);321–328.

[3] De Cordova MIP, Mier N, Quirarte NHG, et al. Role and working conditions of nurses in public health in Mexico and Peru: a binational qualitative study. J Nurs Manag. 2013;21:1034–1043.

[4] Cubaka VK, Schriver M, Flinkenflögel M, et al. The evolving role of physicians - don't forget the generalist primary care providers. Int J Heal Policy Manag. 2016;5:605–606.

[5] Kilminster SM, Jolly BC. Effective supervision in clinical practice settings: a literature review. Med Educ. 2000;34:827–840.

[6] Bosch-Capblanch X, Liaqat S, Garner P. Managerial supervision to improve primary health care in low- and middle-income countries. Cochrane Database Syst Rev. 2011 September;(9):CD006413.

[7] Bosch-Capblanch X, Garner P. Primary health care supervision in developing countries. Trop Med Int Heal. 2008 Mar;13:369–383.

[8] Bailey C, Blake C, Schriver M, et al. A systematic review of supportive supervision as a strategy to improve primary healthcare services in Sub-Saharan Africa. Int J Gynecol Obstet. 2016 Nov;132:117–125.

[9] Vasan A, Mabey DC, Chaudhri S, et al. Support and performance improvement for primary health care workers in low- and middle-income countries: a

scoping review of intervention design and methods. Health Policy Plan. 2017;32:437–452.

[10] The United Republic of Tanzania Ministry of Health and Social Welfare. A manual for comprehensive supportive supervision and mentoring on HIV and AIDS health services [Internet]. 2014. Available from: www.nacp.go.tz/site/download/Manual_CSSM_2nd_Edition_2014.pdf

[11] Government of Uganda Ministry of Health. Health Sector Strategic Plan III 2010/11-2014/15 [Internet]. 2010. Available from: http://www.docucu-archive.com/view/3fd03570993e100299701841c04adae9/Health-Sector-Strategic-Plan-III-2010/11-2014/15.pdf

[12] Federal Democratic Republic of Ethiopea Ministry of Health. Health Sector Development Programme IV 2010/11-2014/15 [Internet]. 2010. Available from: https://phe-ethiopia.org/admin/uploads/attachment-721-HSDPIVFinalDraft11Octoberr2010.pdf

[13] Marquez L, Kean L. Making supervision supportive and sustainable: new approaches to old problems. Washington (DC): Maximizing Access Qual Initiat USAID; 2002. MAQ Paper.

[14] Schriver M, Cubaka VK, Itangishaka S, et al. Perceptions on evaluative and formative functions of external supervision of Rwandan primary healthcare facilities. A qualitative study. PLoS One. 2018;13(2).

[15] John Clements C, Streefland PH, Malau C, et al. Supervision in primary health care-can it be carried out effectively in developing countries? Curr Drug Saf. 2007;2:19–23.

[16] Wheeler S, Barkham M. A core evaluation battery for supervision. In: Watkins CE Jr, Milne DL, editors. The Wiley international handbook of clinical supervision. Chichester:John Wiley & Sons; 2014. p. 367–385.

[17] Vonk ME, Thyer BA. Evaluating the quality of supervision: A review of instruments for use in field instruction. Clin Superv. 1997;15:103–113.

[18] Fluit CRMG, Bolhuis S, Grol R, et al. Assessing the quality of clinical teachers: a systematic review of content and quality of questionnaires for assessing clinical teachers. J Gen Intern Med. 2010;25:1337–1345.

[19] Schriver M, Cubaka VK, Nyirazinyoye L, et al. The relationship between primary healthcare providers and their external supervisors in Rwanda. African J Prim Heal Care Fam Med. 2017;9:1.

[20] Hegarty WH. Using subordinate ratings to elicit behavioral changes in supervisors. J Appl Psychol. 1974;59:764–766.

[21] Coltman T, Devinney TM, Midgley DF, et al. Formative versus reflective measurement models: two applications of formative measurement. J Bus Res Elsevier Inc. 2008;61:1250–1262.

[22] Proctor B. Supervision: a co-operative excercise in accountability. In: Marken M, Payne M, editors. Enabling & ensuring supervision in practice. Leicester: National Youth Bureau; 1987. p. 21–34.

[23] Rohde J. Supportive supervision to improve integrated primary health care. Manag Sci Heal Occas Pap. 2006;2:1–44.

[24] WHO, The Department of Immunization V and B. Module 4: supportive supervision [Internet]. Training for mid-level managers series. Geneva: WHO Document Production Services; 2008. Available from: www.who.int/vaccines-documents/

[25] EngenderHealth. Facilitative supervision handbook [Internet]. Engender Health's Quality Improvement Series. New York; 2001. Available from: https://www.engenderhealth.org/pubs/quality/facilitative-supervision-handbook.php

[26] EngenderHealth, USAID. Facilitative supervision for quality improvement: participant handbook [Internet]. The ACQUIRE Project, editor. 2008. Available from: http://www.acquireproject.org/fileadmin/user_upload/ACQUIRE/Facilitative-Supervision/Participants-Handbook/FS_PartHandbk_main_text.pdf

[27] Bhana R, The Health Systems Trust SA. Supportive supervision system for district health management teams : a guide to primary health care supervison [Internet]. 2010. Available from: www.hst.org.za

[28] Management and Leadership Program. Supervision guidelines [Internet]. Available from: https://www.k4health.org/sites/default/files/SupervisionGuidelines_MSH.pdf

[29] Children's Vaccine Program at PATH. Guidelines for Implementing Supportive Supervision: A step-by-step guide with tools to support immunization [Internet]. Seattle:PATH; 2003. Available from: http://www.path.org/vaccineresources/files/Guidelines_for_Supportive_Supervision.pdf

[30] Benavente J, Madden C, editors. Improving supervision: a team approach. The family planning manager: management strategies for improving family planning service delivery. 1993. p. 1–17. Available from: http://www.wahooas.org/mshdvd2/pdf_managers_Eng/HRM_HCD/Imp_Sup_2_5_Eng_issue.pdf

[31] Tavrow P, Kim Y-M, Malianga L. Measuring the quality of supervisor-provider interactions in health care facilities in Zimbabwe. Int J Qual Heal Care. 2002 Dec;14:57–66.

[32] Beaton DE, Bombardier C, Guillemin F, et al. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine (Phila Pa 1976). 2000;25:3186–3191.

[33] García AA. Cognitive interviews to test and refine questionnaires. Public Health Nurs. 2011;28:444–450.

[34] de Vet HCW, Terwee CB, Mokkink LB, et al. Measurement in medicine: a practical guide. Cambridge, UK: Cambridge University Press; 2011. p. 338.

[35] Costello AB, Osborne JW. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. Pract Assessment Res Eval. 2005;10:1–9.

[36] Hair JF, Black WC, Babin BJ, et al. Multivariate data analysis. 7th ed. Essex: Pearson Educated Limited; 2014.

[37] Song J, Belin TR. Choosing an appropriate number of factors in factor analysis with incomplete data. Comput Stat Data Anal. 2008;52:3560–3569.

[38] Williams B, Brown T, Onsman A. Exploratory factor analysis: a five-step guide for novices. Aust Jounal Paramed. 2010;8:3.

[39] Kottner J, Audige L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. Int J Nurs Stud. 2011;48:661–671.

[40] Vanbelle S. A new interpretation of the weighted kappa coefficients. Psychometrika. 2016;81:399–410.

[41] Polit DF, Beck CT. The content validity index: are you sure you know what ' s being reported? critique

and recommendations. Res Nurs Heal. 2006;29:489–497.

[42] Polit DF, Beck CT, Owen SV. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Res Nurs Health. 2007;30:459–467.

[43] Bello DA, Hassan ZI, Afolaranmi TO, et al. Supportive supervision: an effective intervention in achieving high quality malaria case management at primary health care level in Jos, Nigeria. Ann Afr Med. 2013;12:243–251.

[44] Langston A, Weiss J, Landegger J, et al. Plausible role for CHW peer support groups in increasing care-seeking in an integrated community case management project in Rwanda: a mixed methods evaluation. Glob Heal Sci Pract. 2014;2:342–354.

[45] Curran PJ, West SG, Finch JF. The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. Psychol Methods. 1996;1:16–29.

[46] Satorra A, Bentler PM. A scaled difference chi-square test statistic for moment structure analysis. Psychometrika. 2001;66:507–514.

[47] McDonald RP, Ho M-HR. Principles and practice in reporting structural equation analyses. Psychol Methods. 2002;7:64–82.

[48] Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Model A Multidiscip J. 1999;6:1–55.

[49] Hoyle RH. Handbook of structural equation modeling. New York, NY: Guilford Press; 2012. p. 740.

[50] UCAR Finance and administration. Employee evaluation of supervision - form [Internet]. 2001. Available from: http://www.fin.ucar.edu/forms/HR/supereval_form/supereval_form.shtml

[51] Holloway EL, Wampold BE, University of Utah. Dimensions of satisfaction in the supervision interview. In: Annual convention of the American Psychological Association (92nd); 1984 Aug 24–28; Toronto, ON.

[52] South Africa National Department of Health. 4.6 Checklist: PHC supervisor/supervisee relationship. In: Primary health care supervision manual: a guide to primary health care facility supervision. Pretoria; 2009.

[53] Oklahoma State University. Supervisory staff performance evaluation [Internet]. Available from: https://hr.okstate.edu/sites/default/files/docfiles/evaluation_form_supervisors.doc

[54] Shapiro M, Pogosjana M, Rylander A, et al. What about supervision matters? - The influence of supervision satisfaction on turnover intentions. In: Annual Convention of Association of Behavior Analysis International; 2011; Denver, CO. Available from: https://klab-csun.weebly.com/conference-presentations.htm

[55] Ladany N, Hill CE, Corbett MM, et al. Nature, extent, and importance of what psychotherapy trainees do not disclose to their supervisors. J Couns Psychol Nondisclosure. 1996;43:10–24.

[56] Löfmark A, Thorkildsen K, Råholm MB, et al. Nursing students' satisfaction with supervision from preceptors and teachers during clinical practice. Nurse Educ Pract. 2012;12:164–169.

[57] Noe RA. An investigation of the determinants of successful assigned mentoring relationships. Pers Psychol. 1988;41:457–479.

[58] O'Donovan A, Halford WK, Walters B. Towards best practice supervision of clinical psychology trainees. Aust Psychol. 2011;46:101–112.

[59] Suen LKP, Chow FLW. Students' perceptions of the effectiveness of mentors in an undergraduate nursing programme in Hong Kong. J Adv Nurs. 2001;36:505–511.

[60] Uys LR, Minnaar A, Reid S, et al. The perceptions of nurses in a district health system in KwaZulu-Natal of their supervision, self-esteem and job satisfaction. Curationis. 2004;27:50–56.

[61] Boerboom T, Dolmans D, Jaarsma A, et al. Exploring the validity and reliability of a questionnaire for evaluating veterinary clinical teachers' supervisory skills during clinical rotations. Med Teach. 2011;33:e84–91.

[62] Dilmore TC, Rubio DM, Cohen E, et al. Psychometric properties of the mentor role instrument when used in an academic medicine setting. Clin Transl Sci. 2010;3:104–108.

[63] Fleming M, House S, Hanson VS, et al. The mentoring competency assessment. Acad Med. 2013;88:1002–1008.

[64] Horton S, de Lourdes Drachler M, Fuller A, et al. Development and preliminary validation of a measure for assessing staff perspectives on the quality of clinical group supervision. Int J Lang Commun Disord. 2008;43:126–134.

[65] McGilton KS. Development and psychometric testing of the supportive supervisory scale. J Nurs Sch. 2010 July 14;42:223–232.

[66] Palomo M, Beinart H, Cooper MJ. Development and validation of the Supervisory Relationship Questionnaire (SRQ) in UK trainee clinical psychologists. Br J Clin Psychol. 2010 Jun;49:131–149.

[67] Rogers J, Monteiro FM, Nora A. Toward measuring the domains of mentoring. Fam Med. 2008;40:259–263.

[68] Seki Sakakibara K, Ishikawa H, Kiuchi T. Reliability and validity of the Japanese version of the mentoring functions questionnaire 9-item version. Sangyo Eiseigaku Zasshi. 2013;55:125–134.

[69] Stebnicki MA, Allen HA, Janikowski TP. Development of an instrument to assess perceived helpfulness of clinical supervisory behaviours. Rehabil Educ. 1997;11:307–322.

[70] Winstanley J, White E. The MCSS-26: revision of the Manchester clinical supervision scale; using the rasch measurement model. J Nurs Meas. 2011;19:160–178.

[71] Zarbock G, Drews M, Bodansky A, et al. The evaluation of supervision: construction of brief questionnaires for the supervisor and the supervisee. Psychother Res. 2009;19:194–204.

[72] Makoul G, Krupat E, Chang CH. Measuring patient views of physician communication skills: development and testing of the communication assessment tool. Patient Educ Couns. 2007;67:333–342.

[73] Mercer SW, Maxwell M, Heaney D, et al. The consultation and relational empathy (CARE) measure: development and preliminary validation and reliability of an empathy-based consultation process measure. Fam Pract. 2004;21:699–705.

[74] Bresick G, Sayed AR, Le Grange C, et al. Adaptation and cross-cultural validation of the United States primary care assessment tool (expanded version) for use in South Africa. African J Prim Heal Care Fam Med. 2015;7:1–11.

[75] Kent P, Lauridsen HH. Managing missing scores on the roland morris disability questionnaire. Spine (Phila Pa 1976). 2011;36:1878–1884.

[76] Stalmeijer RE, Dolmans DHJM, Wolfhagen IH, et al. The Maastricht Clinical Teaching Questionnaire (MCTQ) as a valid and reliable instrument for the evaluation of clinical teachers. Acad Med. 2010;85:1732–1738.

[77] Martinko MJ, Harvey P, Brees JR, et al. A review of abusive supervision research. J Organ Behav. 2013;34:120–137.

[78] Brees J, Martinko M, Harvey P. Abusive supervision: subordinate personality or supervisor behavior? J Manag Psychol. 2016;31:405–419.

[79] LaVela S, Gallan A. Evaluation and measurement of patient experience. Patient Exp J. 2014;1:28–36.

[80] Haddad S, Potvin L, Roberge D, et al. Patient perception of quality following a visit to a doctor in a primary care unit. Fam Pract. 2000;17:21–29.