We-Act Measure: Feasibility and Sensitivity to Change Testing among Service Providers in Gaza

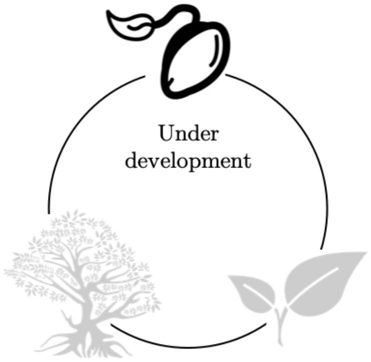
Mark Jordans[1,2], April Coetzee[2], & Frederik Steen[2]
[1] Kings College London
[2] War Child Holland

Technical Working Paper
Draft August 2019

Abstract

The We-Act competency measurement tool has been developed to measure 14 competencies that are seen as foundational for service providers across sectors that work directly with children. These include group facilitators, teachers and educators. The study outlined in this report tested the tool in Gaza with experienced and new facilitators of a low level, psychosocial support interventions, the DEALS. The mixed-method study tested the tool on feasibility, reliability and sensitivity to change in competencies as well as the interrater reliability of the competency raters. Competencies were rated through structured live role plays that provided the opportunity for the facilitators to show competency levels in all 14 competencies. Two groups of facilitators were rated, facilitators who were experienced in the DEALS methodology and facilitators new to it who were rated pre and post training. Preliminary findings show tool seems to be working and feasible, the psychometric properties are adequate, the tool is usable and the format good, and the tool is able to capture the change that we expected after training.

Overview of We-Act: MENAT Measurement Library Criteria



We-Act should have high evidence of reliability, validity, and specificity/sensitivity for use as a screening tool; moderate to high evidence of validity, reliability, and sensitivity to change for use as an evaluation measure; and low to moderate evidence of validity and reliability for use as supervision tool with the goal of providing feedback to service providers. There is not yet evidence of the types of reliability and validity needed for use of the measure for its specified purposes. Given the sampling design, sample size, and rigor of psychometric method, there is uncertainty in the replicability and/or accuracy of the evidence if tested with a similar sample. If interested in use of the measure, please contact the developer for further information.

Criteria	Indicators	Notes
Purposes	Program/training evaluation and research	Requires high interrater reliability; strong evidence of validity; sensitivity to change
	Monitoring and supervision of staff for improvement purposes	Requires less stringent criteria
	Screening of service provider candidates	Requires evidence of sensitivity/specificity
Empirical evidence overall	# of types of evidence available	3
	% of evidence meets empirical criteria	48% (green only); 50% (yellow and green)
	Evidence fit for purpose	Yes for sensitivity to change and promising for interrater reliability; not yet for validity or sensitivity/specificity
Confidence in evidence	Sampling method	Convenience sampling
	Sample size	Small
	Missing data	Small amount of missing data
	Rigor of method	Moderate
Revisions	Clear guidance on what to adjust/refine	No

Overview of We-Act Empirical Results

Competencies/items assessed	Feasibility	Inter-rater reliability	Sensitivity to change	Recommendations for revision
Overall competency	✓	○	○	
Non-verbal communication	✓	NE	✓	
Verbal communication	✓	NE	✓	
Rapport	✓	NE	✓	
Empathy	✓	NE	✓	
Supports reflection and reframing	✓	NE	NE	
Facilitates group work	✓	NE	NE	
Ensures meaningful participation	✓	NE	○	
Demonstrates behavior management	✓	NE	NE	
Demonstrates problem solving	✓	NE	NE	
Identifies child's needs	✓	NE	NE	
Detects child abuse	✓	NE	✓	
Demonstrates collaboration	✓	NE	NE	
Ability to be inclusive	✓	NE	✓	
Give and receive feedback	✓	NE	NE	

Key

✓	Good/excellent evidence against empirical criteria	○	Fair/inconclusive evidence against empirical criteria	✗	Little to no evidence against empirical criteria at this juncture[1]	**NA** **NE**	Not applicable Not evaluable

For additional information on the empirical criteria, please see https://inee.org/measurement-library

This technical working paper was developed by Mark Jordans, April Coetzee, and Frederik Steen as members of the 3EA | MENAT Measurement Consortium, and reviewed by NYU Global Ties for Children.

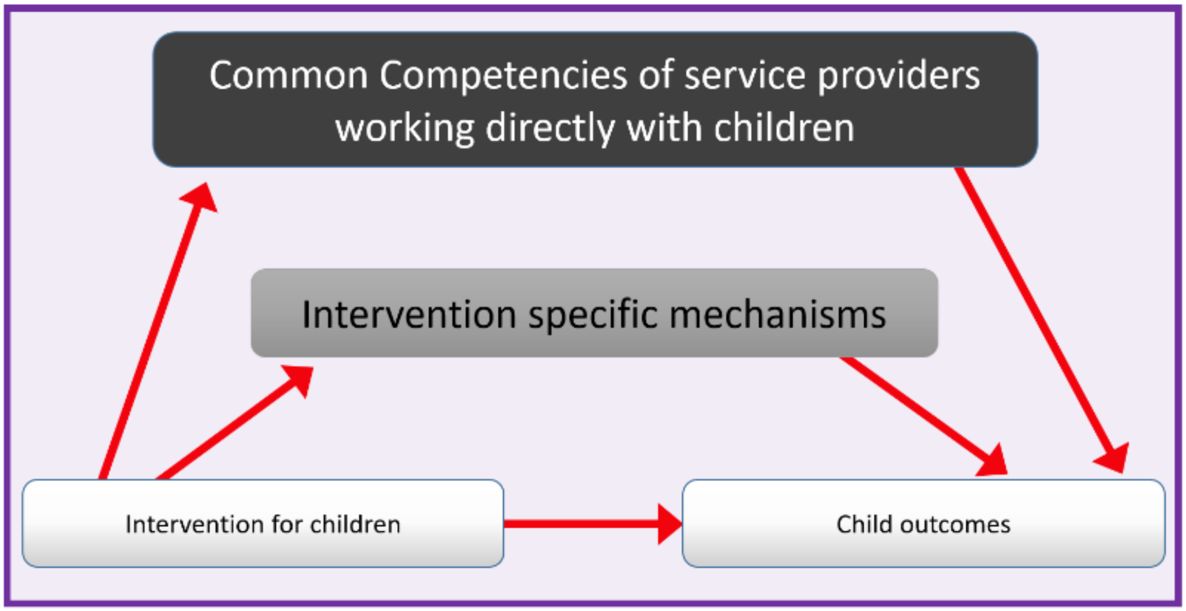
Suggested citation: Jordans, M., Coetzee, A., & Steen, F. (2019, August). *We-Act measure: Feasibility and sensitivity to change testing among service providers in Gaza.* Technical working paper. Amsterdam: War Child Holland.

Introduction

Problem statement

- How do we improve the quality of care with children in humanitarian and low resource settings, when working with non-specialists across different sectors?
- How do we adapt training programs to target key gaps in competencies of service providers?
- How do we have a systematic approach to supervision with a clear framework towards mastery of competencies?

Figure 1. *Relationship between measurement of competencies and quality outcomes*

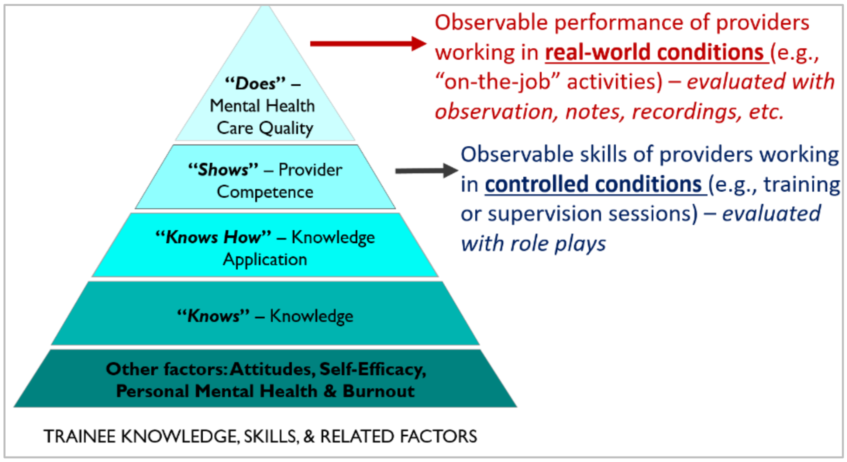

Common Competencies of service providers working directly with children
Intervention specific mechanisms
Intervention for children
Child outcomes

Background

The aim of the We-Act measure is to assess a set of common competencies across an integrated care system for frontline workers, facilitators and teachers in education settings in fragile and conflict affected states that have face-to-face contact with children. Competence is defined as the extent to which a frontline worker has the knowledge and skill required to deliver an intervention to the standard needed for it to achieve its expected effects.

The measure is based on based on George Miller's framework (Miller, 1990) for (clinical) competence. This framework distinguishes between what a person knows (knowledge), knows how (knowledge application), shows how (competency), and does (performance in real world conditions). The measurement tool will take this into account through the introduction of a scoring system from non-expert to expert for each competence, with the recognition that competence is reached when frontline workers can demonstrate competence beyond knowledge to application in their day to day practice.

Figure 2. *George Miller framework*


"Does" – Mental Health Care Quality
"Shows" – Provider Competence
"Knows How" – Knowledge Application
"Knows" – Knowledge
Other factors: Attitudes, Self-Efficacy, Personal Mental Health & Burnout
TRAINEE KNOWLEDGE, SKILLS, & RELATED FACTORS
Observable performance of providers working in real-world conditions (e.g., "on-the-job" activities) – evaluated with observation, notes, recordings, etc.
Observable skills of providers working in controlled conditions (e.g., training or supervision sessions) – evaluated with role plays

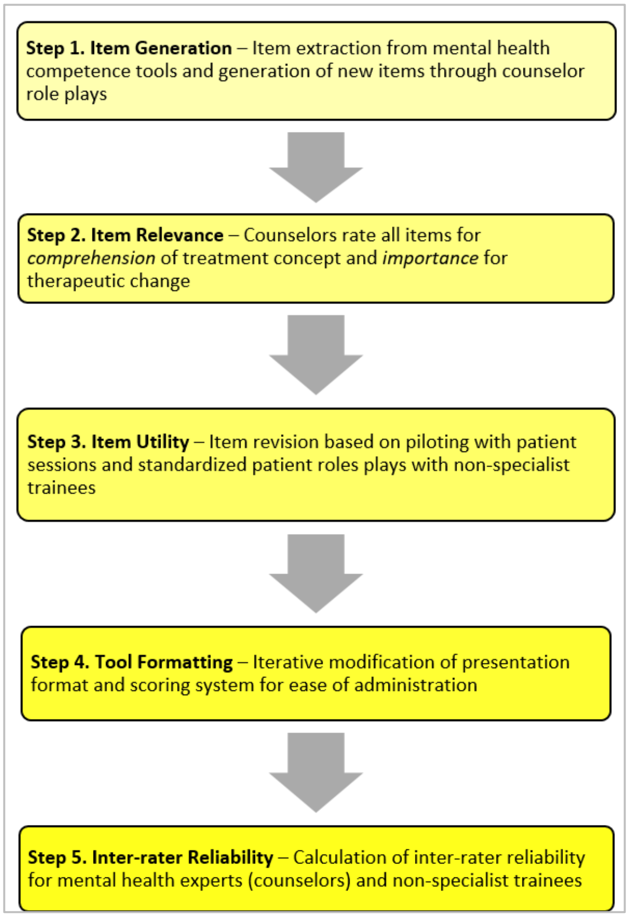
Tool development

The development of the tool draws inspiration from similar work that has already been done in Nepal for service providers in the Mental Health & Psychosocial Support sector working with adults. The Enhancing Assessment of Common Therapeutic Factors (ENACT) is an observational rating scale developed for interventions in which non-specialists are trained to deliver mental health and psychosocial interventions. (Luitel et al., 2015; Ramaiya, Bhardwaj, Rai, Kohrt, & Jordans, 2015; Subba, Luitel, Kohrt, & Jordans, 2017)

The following stages were followed to identify 14 competencies in final We-Act tool:

i. Through six sources, items were collected that looked at competencies across education, child protection and mental health.

Figure 3. *ENACT tool development process*


Step 1. Item Generation – Item extraction from mental health competence tools and generation of new items through counselor role plays
Step 2. Item Relevance – Counselors rate all items for comprehension of treatment concept and importance for therapeutic change
Step 3. Item Utility – Item revision based on piloting with patient sessions and standardized patient roles plays with non-specialist trainees
Step 4. Tool Formatting – Iterative modification of presentation format and scoring system for ease of administration
Step 5. Inter-rater Reliability – Calculation of inter-rater reliability for mental health experts (counselors) and non-specialist trainees

ii. Items were sorted, grouped and prioritised until a final selection of 14 competencies was agreed upon.

iii. Final process to operationalise competencies through 4-point scale from potential harm, absence of competency, competent and mastery.

Figure 4. *Competency identification process*

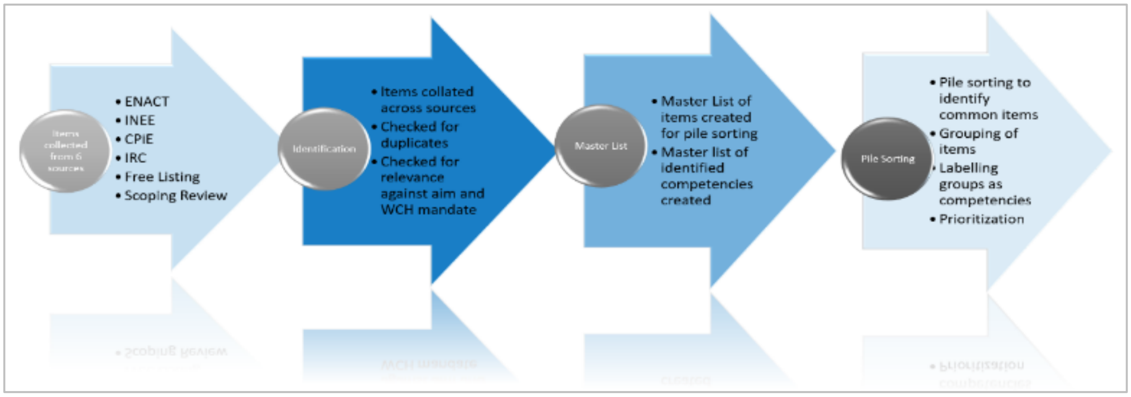
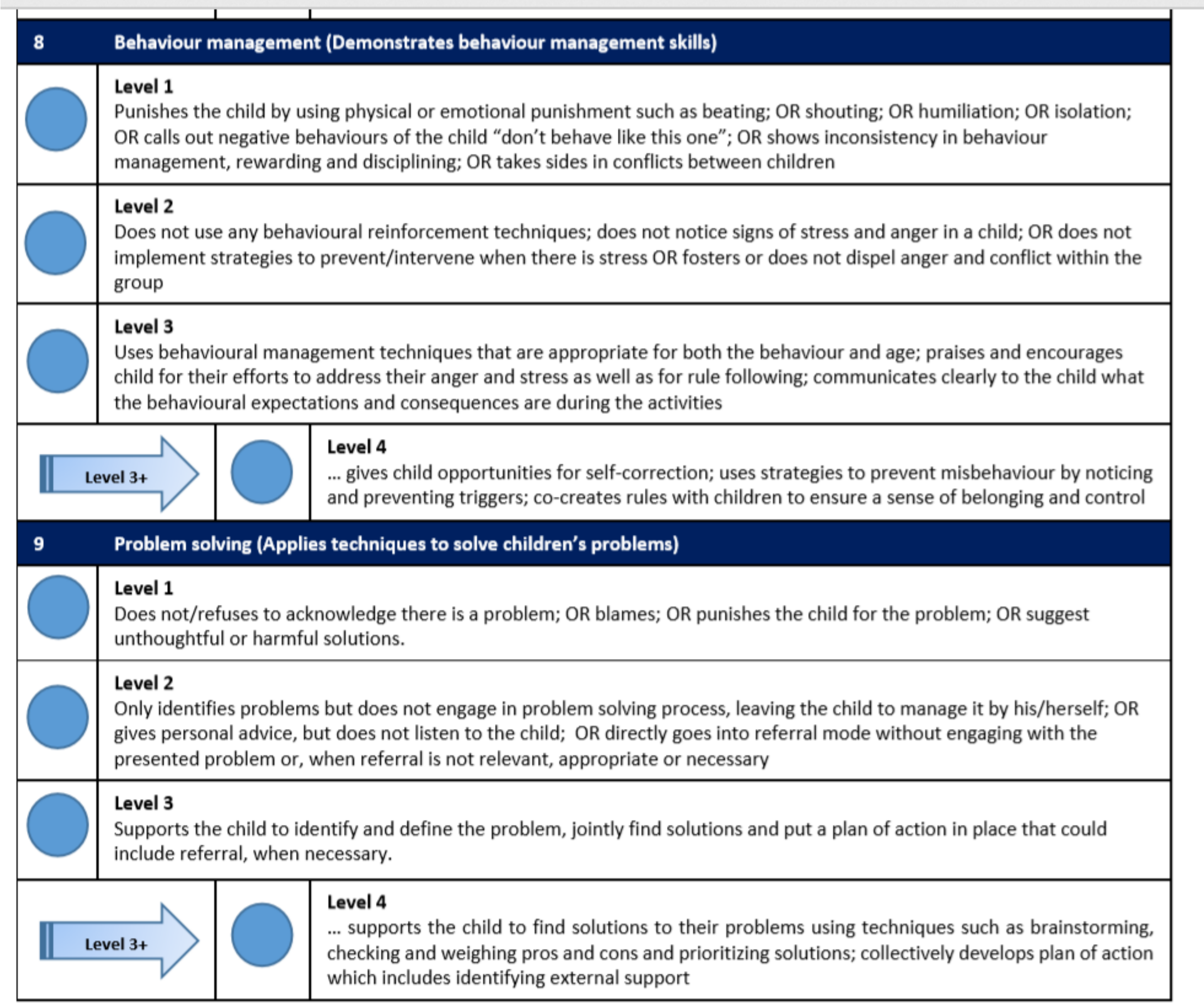

Items collected from 6 sources
• ENACT
• INEE
• CPiE
• IRC
• Free Listing
• Scoping Review
Identification
• Items collated across sources
• Checked for duplicates
• Checked for relevance against aim and WCH mandate
Master List
• Master List of items created for pile sorting
• Master list of identified competencies created
Pile Sorting
• Pile sorting to identify common items
• Grouping of items
• Labelling groups as competencies
• Prioritization

Table 1. *List of We-Act competencies*

1. Non-verbal communication
2. Verbal communication
3. Rapport and relationship building
4. Empathy, warmth and genuineness
5. Support the Reflection and Reframing of Children's Thoughts and Feelings
6. Manages and facilitates group work effectively
7. Ensures children's meaningful participation
8. Demonstrates behaviour management skills
9. Applies problem solving techniques for the child's problems
10. Ability to identify and understand child's needs
11. Detect and observe for child abuse, exploitation, neglect, violence, and self-harm
12. Demonstrates collaboration with caregivers and other actors
13. Ability to be inclusive
14. Giving and receiving feedback

Figure 5. *Example of We-Act competencies and levels*



8	**Behaviour management (Demonstrates behaviour management skills)**	
●	**Level 1** Punishes the child by using physical or emotional punishment such as beating; OR shouting; OR humiliation; OR isolation; OR calls out negative behaviours of the child "don't behave like this one"; OR shows inconsistency in behaviour management, rewarding and disciplining; OR takes sides in conflicts between children	
●	**Level 2** Does not use any behavioural reinforcement techniques; does not notice signs of stress and anger in a child; OR does not implement strategies to prevent/intervene when there is stress OR fosters or does not dispel anger and conflict within the group	
●	**Level 3** Uses behavioural management techniques that are appropriate for both the behaviour and age; praises and encourages child for their efforts to address their anger and stress as well as for rule following; communicates clearly to the child what the behavioural expectations and consequences are during the activities	
Level 3+	●	**Level 4** … gives child opportunities for self-correction; uses strategies to prevent misbehaviour by noticing and preventing triggers; co-creates rules with children to ensure a sense of belonging and control
9	**Problem solving (Applies techniques to solve children's problems)**	
●	**Level 1** Does not/refuses to acknowledge there is a problem; OR blames; OR punishes the child for the problem; OR suggest unthoughtful or harmful solutions.	
●	**Level 2** Only identifies problems but does not engage in problem solving process, leaving the child to manage it by his/herself; OR gives personal advice, but does not listen to the child; OR directly goes into referral mode without engaging with the presented problem or, when referral is not relevant, appropriate or necessary	
●	**Level 3** Supports the child to identify and define the problem, jointly find solutions and put a plan of action in place that could include referral, when necessary.	
Level 3+	●	**Level 4** … supports the child to find solutions to their problems using techniques such as brainstorming, checking and weighing pros and cons and prioritizing solutions; collectively develops plan of action which includes identifying external support

Research Questions

The study evaluated the psychometric properties of the We-Act tool, specifically: (a) feasibility; (b) reliability; and (c) sensitivity to change. We used a mixed methods study design, including focus group discussions for (a), assessing interrater reliability for (b), and testing of pre-post change scores for (c). More specifically:

a. Can expert raters and field staff use We-Act tool to rate common competencies?
b. Is We-Act tool reliable, so that independent expert raters using it come to similar scores after observing a service provider through standardized role plays?
c. Is there a change in We-Act tool scores over time and training?

Sample

The research sample consisted of service providers experienced in the DEALS methodology (*N*=64) and those who are new to it (*N*=25). Data was also collected from competency expert raters (*N*=8).

Table 2. *Demographics of facilitators*

	New Facilitators	**Experienced Facilitators**
Total number facilitators	25	64
Gender	19 Female, 6 Male	41 Female, 23 Male
Average age	24.56	28.42
Years of Psychosocial experience	0.88	4.98
Years of facilitating DEALS	0	1.81

Table 3. *Demographics of competency raters*

	Competency Raters
Total number raters	8
Gender	6 Female, 2 Male
Average age	29.5
Years of supervision experience	1.5
Education Background	maths, english, counselling, arabic, science

Convenience sampling procedure followed with service providers being selected from within ongoing War Child and partner community-based organisation. Competency raters were selected **internally**, from existing supervisors of ongoing DEALS programmes within War Child and **externally**, from local CBOs that adopted the DEALS methodology within their own programming. This ensured that all competency raters had a similar background but supported the external validity of the psychometric evidence.

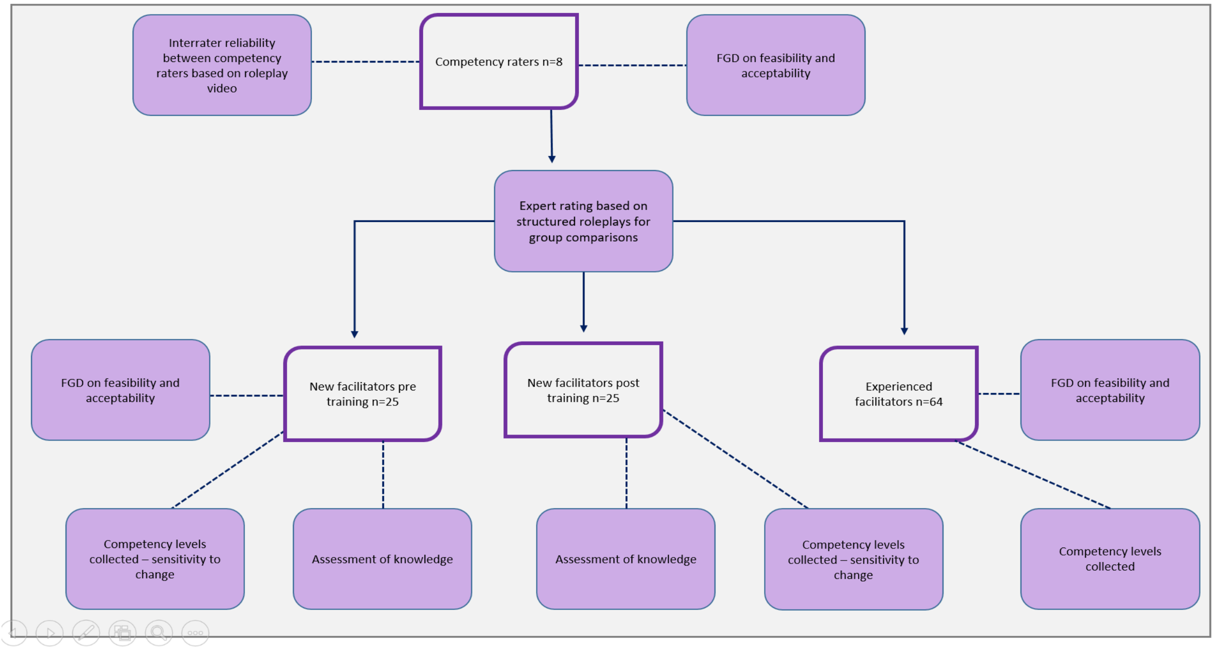
Methods

Development of competency role-plays

Competencies were tested through standardised role plays that were performed live when rating the facilitators. Scenarios were developed and scripted with a local theatre group to ensure consistency. Each competency was represented to ensure that the facilitators would have an opportunity to demonstrate each competency. Structured reactions were also developed in order to respond, in real time, to the actions of the facilitators. The competency raters were assigned to each live rating session rather than randomised to ensure that they were rating facilitators from outside their organisation, therefore reducing the possible effect of bias and over rating through familiarisation. The same three raters rated the pre- and post-training roleplays. In the development of the roleplays:

a. Two scenarios were developed: A group scene showing 9 competencies and an individual scene showing 5 competencies.
b. Scripts developed in partnership with local theatre group to ensure a mixture of levels across the competencies in the scenarios.
c. Standardised role plays simulated real-life scenarios that the service providers experienced when implementing the programme.
d. Standardised role plays recorded on video for interrater reliability (IRR) and performed live for rating of competencies.

Figure 6. *Overview of study plan for We-Act*


Interrater reliability between competency raters based on roleplay video
Competency raters n=8
FGD on feasibility and acceptability
Expert rating based on structured roleplays for group comparisons
FGD on feasibility and acceptability
New facilitators pre training n=25
New facilitators post training n=25
Experienced facilitators n=64
FGD on feasibility and acceptability
Competency levels collected – sensitivity to change
Assessment of knowledge
Assessment of knowledge
Competency levels collected – sensitivity to change
Competency levels collected

Training of competency raters

A 3-day training was held with competency raters (*N*=8). Training included:

- understanding of We-Act tool and familiarity in use of it
- building of common understanding of terms, concepts, terminology and language
- observation skills and rating using the tool
- cognitive bias and mitigating factors

Initial training culminated in calculation of interrater reliability scores to ensure the raters had a shared understanding of the competencies and the measurement levels. This was calculated by the raters watching a video of the same roleplay that would be used in the live rating. The video was scripted to ensure that the 'facilitator' in the video demonstrated a variety of competency levels for the raters to identify.

All eight raters who took part in the competency rating training watched the video at the same time and rated independently using the We-Act tool. These tools were then collected and used to calculate the interrater reliability.

Following analysis of the IRR it was discovered that there was some shared misunderstanding of certain competencies so, after the live ratings of the experienced facilitators, the raters were brought together for a refresher one day training. The training focused on ensuring clearer understanding of these specific competencies and the different levels. Another rater did not attend the recap training, because of logistical circumstances. A second IRR with the remaining 6 raters was then taken using the same video as the first round.

Analytic Plan

Quantitative

During the study the following five types of quantitative data were collected:

(1) Competency rating of experienced DEALS service providers through live performance of standardised roleplay. *(feasibility)*
(2) Pre- and post-training competency rating of new DEALS service providers using standardised role play. *(feasibility and sensitivity to change)*
(3) Pre- and post-training knowledge surveys
(4) Expert rating of video-taped standardised role play for IRR analysis.

These were analysed in the following way:

IRR analyses. The intra-class correlation coefficient (ICC) decomposing variance due to raters (trainer/supervisor) and items was established through rating a videotaped session (ICC (3,1) = y (95% CI); Shrout & Fleiss, 1979). This calculation allows us to understand the extent to which ratings were consistent among raters across all items for this particular observation. It does not, however, provide evidence on consistency of ratings of each item, or evidence on the extent to which ratings were consistent among raters across different observations (e.g., with different service providers or videos). Said otherwise, we cannot know the precision of the ICC estimate: High or low consistency across raters on items could be due to something specific about that observation, and therefore may not be generalizable to other service providers/videos. For the interpretation (with caution) we assumed < .50 is poor, between .50 and .75 is moderate, between .75 and .90 is good and > .90 is excellent (Koo & Li, 2016).

Expert rating of pre- and post-training role-plays by new service providers. We summarized the scores on the We-Act of all observed roleplays (pre-training, post-training), analysed as

(1) as a % scoring >3,

(2) as a total score.

(3) the difference between pre- and post-training scores for observed roleplays (n=20), through t-test of total scores and a Chi-square test of proportions.

Pre- and post-training knowledge surveys based on knowledge gained during the DEALS training which was not specifically focused on training on the We-Act competencies. We assessed the correlation, using a Pearson-r analyses, between knowledge total score and We-Act scores at pre- and post-training points.

Expert ratings of role-plays by experienced service providers. We analysed the Cronbach's alpha based on all expert ratings of roles plays with experienced service providers. We also summarized the scores on the We-Act of all observed roleplays as a % scoring >3, as well as a total score.

Qualitative

Qualitative data was collected through 5 focus group discussions with competency raters (*N*=7), experienced facilitators (*N*=20) and new service providers (*N*=10). Through focus group discussions, we assessed the perspectives of users and other relevant stakeholders on the feasibility of We-Act tool as well as common understanding of concepts, terminology and language. We followed framework analysis using NVivo 11.0 software for analysis of data. A codebook was developed through both inductive and deductive themes by research team. These are still being analysed for themes and findings.

Results

IRR

IRR was measured in two time points:

(1) Time point one - taken after three-day training for all the competency rates (*N* = 8). It was decided to exclude one of the raters who rated significantly differently from the majority of other competency raters on most competencies.

(2) Time point two – taken after the refresher training. One more rater was excluded as they were not able to attend the refresher training. IRR increased for the competencies that were shown in the individual role play and down for the competencies shown in the group role play. After initial IRR it was clear that the greatest difference between the raters was on these, individual roleplay competencies. Therefore, the refresher training focused primarily on 3 competencies, which were all part of the individual roleplay, which may have resulted in an increased IRR for these. The fact that no increase is shown for the 9 group competencies, could be related to the reduced focus on these in the recap training. However, overall IRR increased after the refresher

training. The technical appendix available upon request from the developer provides more detail on the IRR results. We note caution in extrapolating the results beyond this sample, however, given the small sample size of videos on which IRR was assessed.

Rating of the competencies

Our assumption is when using the tool for any purpose, recruitment, training of supervision, that competency has been reached at level 3 or 4 (competent and mastery). Graph 1 shows the changes in the totally number of facilitators showing a 3 or 4 over all competencies before and after the DEALS training. This shows some positive results in terms of the tools ability to measure changes in competencies pre and post training. The overall percentage of experienced facilitators reaching levels 3 and 4 is shown as a comparison to the overall final percentage, post training of the new facilitators.

Graphs 2 and 3 show the same findings, broken down by individual competencies, for new and experienced facilitators respectively.

Figure 7. *% of facilitators rated 3&4 on competency levels*

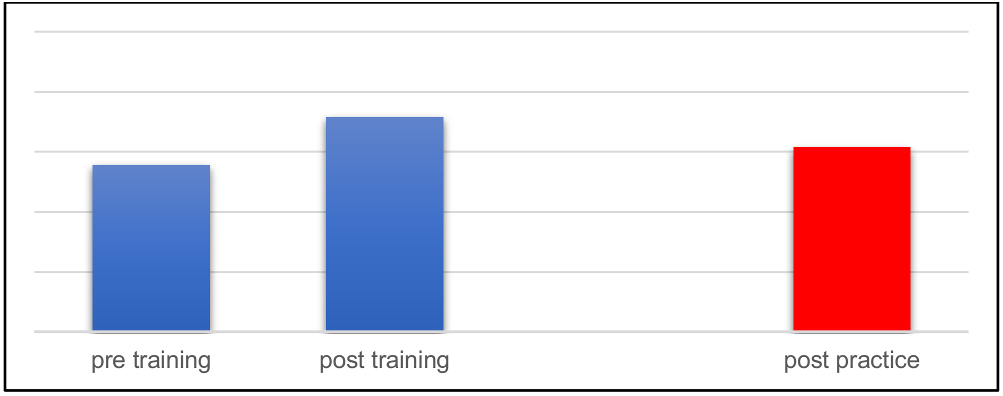


Figure 8. *% of new facilitators scoring 3&4 per competency item*

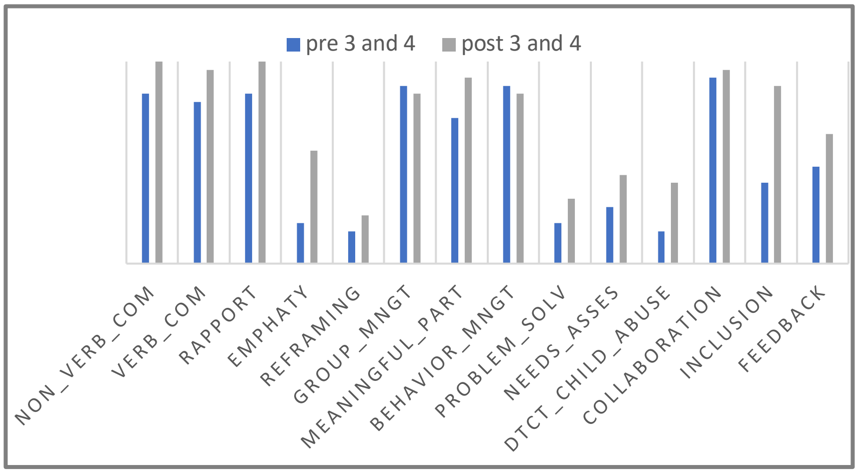
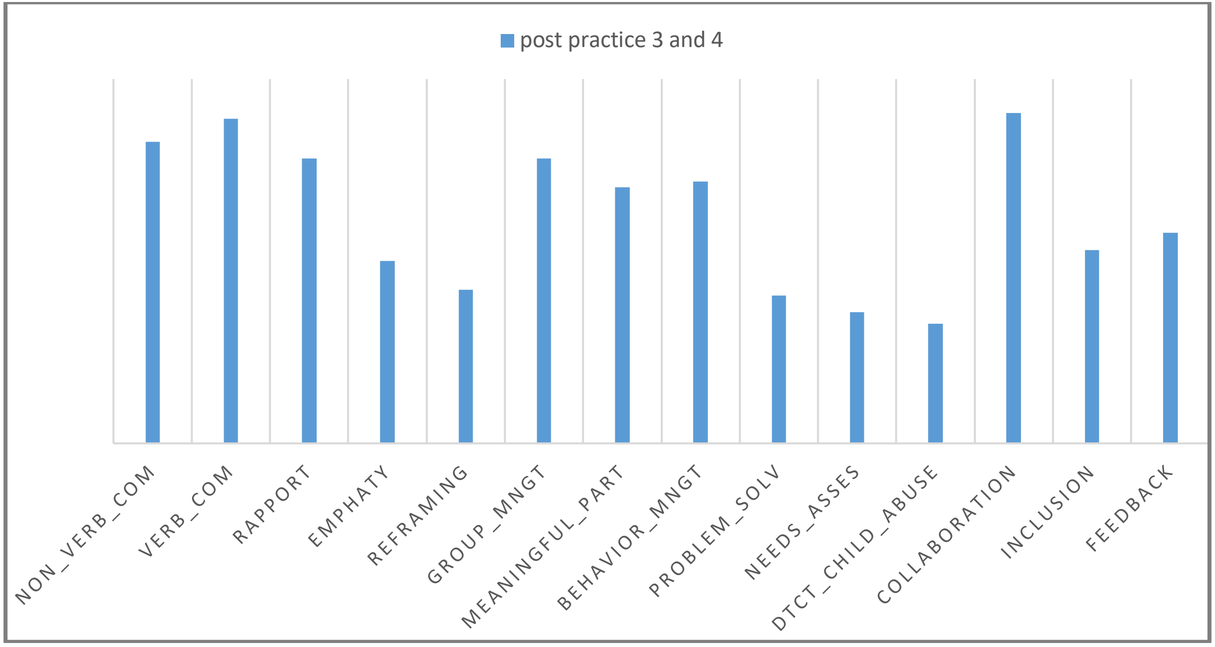


Figure 9. *% of experienced facilitators that scored 3&4 per competency*



From this data we have a number of preliminary findings.

- Tool seems to be working and is feasible with some minor recommendations for change
- Utility of the format is good with the structure of the tool supporting its usability.
- Initial calculation of the consistency of raters across items in the We-Act measure is adequate, but should be replicated across different observations/service providers with a larger sample size to ensure the precision of the estimates. For guidance, see Zou, 2012.
- Whilst the data indicates that the tool is able to capture the change in competency pre- and post-training we recognise that there may be other reasons for this. As the same role play was used both pre- and post- training to measure competencies then there is also a risk of practice effect for both the trainers and raters leading to an increase in competency levels.

Limitations

The competencies of the facilitators were measured using live structured roleplays as described above. During the study a number of lessons were learnt:

- The script used for the IRR video were developed to show a variety of competency levels. During the final video performance there was not clarity in the variance of the levels to be shown in the performance.

- It was a challenge for theatre group to understand the necessity of standardisation and the need to repeat the role play in the same way for each performance, not embellishing or building a theatre piece.
- A language and sector specialist needed to support theatre development to ensure the acting reflects the levels they are demonstrating for the different competencies.
- Ensuring ongoing consistency of theatre group in live performances was challenging
- For this study the decision was made to use adult actors to represent children for a number of child safeguarding and logistical issues. This needs to be a decision for each context reviewing the pros and cons for using adult or child (youth) actors.

Reflection on the competency rater training resulted in the following lessons to take forward for future training:

- The training would benefit from extra video examples of the different levels, especially for some of the more complex competencies such as problem solving and reflection and reframing.
- It was clear during observation of competency raters' first live ratings that they had not a point of reference and therefore overrated the first facilitators, creating something like a ceiling effect of levels. We would recommend that live rating is included as part of the competency training and is done before a recap training day and IRR rating.
- During the training there were a number of points to focus on with the raters:
    - Rating is based on observation and what is seen, rather than assumption. Raters should note down what they observe and then at the end of the session use their notes to rate using the tool.
    - Most raters had experience with observation and supervision. Multiple practice time is needed to ensure that the raters use their experience but are rating against the descriptions in the tool, using the levels and separating out the various competencies (some of which are interconnected).
    - Develop multiple IRR videos in case IRR is poor and retraining is needed and measuring of IRR needs to be repeated, in order to avoid a training effect on IRR rating.
    - Identification of competencies that may be new concepts to the raters to ensure clear understanding,

Other lessons learnt:

- Initially the study was going to explore using peer raters but it was agreed that this was not feasible for this study but will be explored in future studies if possible.
- New facilitators already had almost a year of experience of facilitating PSS workshops. The level of change in competencies pre- and post-training may not have been as large as expected.

Discussion

Moving forward this study has raised a number of points of discussion to consider for future studies and to improve the viability and utility of the tool.

First, how can a golden standard be identified to be able to ensure that competency raters are not only measuring competencies but are rating at the correct level? One possible way of addressing this may be to ensure greater clarity in the level demonstrated in the IRR video which can then be rated by an experienced competency rater or specialist. This can then be used as a golden standard to match against the rating of the video by the trainee competency raters.

Second, at the moment, use of theatre group is the most reliable way to rate 'shows how.' All other ways focus on 'knows how.' What does this mean in terms of scaling up and utilisation of the tool with other organisations, governments etc.? This will be an ongoing discussion for future studies and also as we use the tool in programmes and contexts.

Third, how can the tool be adapted to be able to be used as a supervision tool to measure competencies in real life settings? This study focused on a structured and one off approach to measuring the competencies of the facilitators. Ongoing discussions will explore how the tool can be adapted and used to provide a structured approach to ongoing supervision and providing a structure for the frontline service providers to improve competency levels.

Use of tool

Following the study we would recommend the following points to consider when using this tool:

- Review learnings from competency raters training to ensure competency raters are well equipped to measure using the tool
- Contextualisation is key to ensure that the spirit of each competency is clear in the context and language. Translation, back translation and re translation is key.
- Competencies in the tool should not be merged to make less items. Each of them is unique but interconnected. They may seem as if they cross over but each of them is distinct.

Next steps

War Child would like to undertake the following studies with the We-Act tool:

- Use of tool in different context

- Study on impact of competency on intervention participants' individual outcomes
- Assess feasibility of peer-rating
- Explore competency levels following standard training vs competency driven training.

References

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Luitel, N. P., Jordans, M. J. D., Kohrt, B. A., Ramaiya, M. K., Rai, S., Shrestha, P., … Patel, V. (2015). Therapist competence in global mental health: Development of the ENhancing Assessment of Common Therapeutic factors (ENACT) rating scale. *Behaviour Research and Therapy*, *69*(June 2016), 11–21. https://doi.org/10.1016/j.brat.2015.03.009

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine*, *65*(9), S63-7. https://doi.org/10.1097/00001888-199009000-00045

Ramaiya, M. K., Bhardwaj, A., Rai, S., Kohrt, B. A., & Jordans, M. J. D. (2015). Development of a scoring system for non-specialist ratings of clinical competence in global mental health: a qualitative process evaluation of the Enhancing Assessment of Common Therapeutic Factors (ENACT) scale. *Global Mental Health*, *2*. https://doi.org/10.1017/gmh.2015.21

Subba, P., Luitel, N. P., Kohrt, B. A., & Jordans, M. J. D. (2017). Improving detection of mental health problems in community settings in Nepal: Development and pilot testing of the community informant detection tool. *Conflict and Health*, *11*(1), 1–11. https://doi.org/10.1186/s13031-017-0132-y