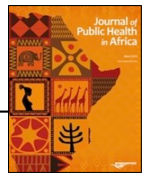
Journal of Public Health in Africa
Received 7 Apr 2022 | Revised xx xxx xxxx | Accepted 15 May 2022 | Published Online 21 Sep 2022

DOI: https://doi.org/10.4081/jphia.2022.2201
JPHIA 13 (3): 2201 (2022)
ISSN (O) 2038-9930 IF:0.0



ORIGINAL ARTICLE

Improving inter-rater reliability of the enhancing assessment of common therapeutic factors (ENACT) measure through training of raters

Mwamba M. Mwenge[1*] | Caleb J. Figge[2] | Kristina Metz[2] | Jeremy C. Kane[3] | Brandon A. Kohrt[4] | Gloria A. Pedersen[4] | Izukanji Sikazwe[1] | Stephanie Skavenski Van Wyk[2] | Saphira M. Mulemba[1] | Laura K. Murray[2]

[1]Centre for Infectious Disease Research in Zambia, Lusaka, Zambia

[2]Johns Hopkins Bloomberg School of Public Health, Baltimore, USA

[3]Columbia University Mailman School of Public Health, New York, USA

[4]George Washington University School of Medicine, Washington, DC, USA, Washington, USA

Abstract

Background. As evidence supports task-shifting approaches to reduce the global mental health treatment gap, counselor competency evaluation measures are critical to ensure evidence-based therapies are administered with quality and fidelity.

Objective. This article describes a training technique for evaluating lay counselors' competency for mental health lay practitioners without rating scale experience.

Methods: Mental health practitioners were trained to give the Enhancing Assessment of Common Therapeutic Factors (ENACT) test to assess counselor proficiency in delivering the Common Elements Treatment Approach (CETA) in-person and over the phone using standardized video and audio recordings. A two-day in-person training was followed by a one-day remote training session. Training includes a review of item scales through didactic instructions, active learning by witnessing and scoring role-plays, peer interactions, and trainer observation and feedback. The trainees rated video and audio recordings, and ICC values were calculated.

Results: The training technique presented in this research helped achieve high counselor competency scores among lay providers with no prior experience using rating scales. ICC rated both trainings satisfactory to exceptional (ICC: .71 - .89).

Conclusions. Raters with no past experience with rating scales can achieve high consistency when rating counselor competency through training. Effective rater training should include didactic learning, practical learning with trainer observation and feedback, and video and audio recordings to assess consistency.

Keywords: training, raters, counselor competency, inter-rater reliability

INTRODUCTION

Common mental health issues, particularly depression and anxiety, are among the primary causes of ill health and account for 7% of the worldwide burden of illness, as measured by disability-adjusted life years (DALYS).[1,2] Seventy percent of the world's population with a mental health issue lacks access to mental health care.

In sub-Saharan Africa, the treatment gap is much greater, ranging from 75% in South Africa to 90% in Ethiopia. Existing and gaining popularity are four effective evidence-based therapies that help bridge the mental health treatment divide. However, a dearth of qualified and competent mental health care professionals, as well as other constraints, prevents the adoption of large-scale treatment provision.[3] Counselor competency is essential to the delivery of mental health services[4,5] because it influences the implementation of evidence-based therapies with quality and fidelity[6] and client outcomes.[7]

The definition of counselor competency can be broken down into two main categories: 1) treatment-specific competence, which focuses on the delivery of a model-specific therapeutic intervention; and 2) global competence or common factors, which focuses on the delivery or demonstration of general therapeutic skills applicable across various treatment models.[8]

There are numerous methods for assessing counselor competency, including essays and questionnaires,[9] competency scales to evaluate typical role-plays, and scoring video or audio recordings of actual therapy sessions.[10]Standardized role-plays preserve standardization and allow raters to watch a counselor's execution of abilities, offering a more realistic evaluation of their proficiency than essays or questionnaires.[11] In order to employ standardized role-plays, a rater must watch and grade the counselor's skill using a competency assessment instrument, and an actor must play the part of a client exhibiting the same mental health disorders in each role-play.[12]

Low inter-rater reliability is one of the difficulties of measuring counselor competency and its relevance to client treatment results.[7]Training raters is a crucial step in enhancing inter-rater reliability; nonetheless, role-plays are frequently evaluated inconsistently by different raters.[13] Therefore, teaching raters appropriately is a crucial step in enhancing inter-rater reliability. Few studies, however, have described the process of teaching competency raters or suggested a training methodology that can assist researchers in achieving high inter-rater reliability.[14]

Current research

We analyzed a training program for counselor competency raters and calculated their inter-rater reliability. This protocol was developed for a research undertaken in Lusaka, Zambia, as part of the Ensuring Quality in Psychosocial Support (EQUIP) project consortium of the World Health Organization (WHO). The WHO EQUIP portal provides a repository of competency assessment instruments, including instructions on how to rate competency, provide competency-based feedback, and perform competency-driven training.[15]

In the EQUIP platform, independent raters assessed counselor competency using the Enhancing Assessment of Common Therapeutic Factors (ENACT)[12] assessment. This study taught counselors in the Common Elements Treatment Approach (CETA).[16] CETA is a flexible, modular, scalable, and evidence-based therapy method designed for lay clinicians to treat co-morbid mental health issues in low- and middle-income nations.[17]

Supplementary information The online version of this article (10.4081/jphia.2022.2201) contains supplementary material, which is available to authorized users.

Corresponding Author: *Mwamba M. Mwenge*
Centre for Infectious Disease Research in Zambia, Lusaka, Zambia. Tel. +26.0976135396.
Email: mmwenge@ymail.com

MATERIALS AND METHODS

Study setting

Two trainings of raters were conducted in Lusaka, Zambia, under the EQUIP Consortium Project. Zambia is located in sub-Saharan Africa and has a population of 17.9 million.[18] Mental health care in Zambia is severely insufficient; service provision is limited to the provincial hospital level and does not currently extend to primary care levels in each district, leading to a significant treatment gap for the wider population.[19]

Participants

Raters and actors

A total of 9 lay providers previously trained in CETA were identified and recruited based on their availability and interest to be trained as competency raters. Providers gave written consent and were trained on how to assess counselor competency using CETA-specific measures. Five providers who had the highest inter-rater reliability during training were selected to be raters for ENACT competency assessments while the remaining 4 were trained to be actors. Characteristics of the 5 raters are summarized in Table 1.

Competency measure

The Enhancing Assessment of Common Therapeutic Factors (ENACT[12]) was originally developed as an 18-item measure of common factors (i.e., general counseling skills) for use in psychological treatment models. The ENACT tool was piloted in Nepal using videotaped role-plays, transcripts from client sessions, and in-person observations of primary care workers in trainings for psychological treatments.[20] The measure has been adapted for use across different cultural settings.[21]

The ENACT tool was adapted for the current study into a 15-item assessment with 4 response options by the EQUIP consortium. Level 1 exhibits unhelpful or potentially harmful behaviours; level 2 demonstrates no basic skills or some but not all skills; level 3 shows all basic skills; and level 4 has all basic skills and any advanced skill. Each level includes a set of observable behaviors that serve to operationalize each of the four levels for each competency item. Adaptations for the Zambia version included modification of a few words (e.g., "helper" was changed to "counselor") to increase cultural appropriateness. The Zambia version of the ENACT was further adapted in September 2020 for use via telephone following the onset of the COVID-19 pandemic and initiation of CETA delivery by telephone.

Study procedure

In total, raters participated in two training sessions (one in-person, one via telephone). Following the in-person training, the 5 raters each observed and evaluated role-plays for 34 counselors at two time points: before the two-week CETA counselor training and immediately after the CETA counselor training. Following the telephone training, the 5 raters also evaluated telephone role-plays for 17 out of the 34 counselors at three time points: pre-telephone counselor training; immediately post-training, and post-supervision. The study procedure flow is summarized in Figure 1.

In-person training

The two-day in-person training was held at the Centre for Infectious Disease Research in Zambia in January 2020. (CIDRZ). The training began with an overview of the EQUIP study and raters' responsibilities. Each rater was given a copy of the ENACT measure for a full item evaluation, and comments and clarifications from the raters were encouraged.

The five raters were then divided into two groups, given several copies of the ENACT assessment, and taught to practice competency rating. In Group 1, one rater acted as both a client and rater, while the other acted as a counselor. In Group 2, which consisted of three raters, one rater acted as a counselor, another as a client, and the third served as the observer and scorer of the role-plays. After two or three role-plays, the trainer switched groups and ensured that each rater practiced scoring while the trainer observed and scored concurrently. The instructor read the actor prompts (i.e., actor difficulties and symptoms) for each role-play to both groups. Each role-play lasted around four to six minutes. The trainer co-rated with trainees and supplied at least three pieces of feedback to each individual rater. The trainer requested the

rating and justification for each item. The trainer compared the rater's scores to the expert's scores, noting difficult-to-score items and rater questions. The instructor resolved conflicts and clarified the reasoning behind the expert scores.

Next, the trainer presented all raters with general input regarding the accuracy of their scores. The instructor encouraged talks among raters in order to establish shared scoring conventions for every item. The instructor facilitated this procedure to verify that all agreed-upon norms were correct and understood by all raters. The trainer encouraged all raters to participate in the conversation and presented a more in-depth reason for counselor behaviors that typically resulted in disparate ratings among raters.

Finally, the trainer devised role-play scenarios based on the most challenging counselor behaviors for raters to effectively and consistently evaluate. Trainers functioned as counselors while evaluating trainer role-plays. The trainer changed the level of proficiency in each role-play and instructed the evaluators to take thorough notes. The trainer requested that the raters debate and justify their scores with one another. The trainer solicited the group's final consensus rating.

Video recordings of counseling sessions

On the second day of rater training, participants viewed four video recordings of an introduction to counseling session. Four movies featuring counselors displaying a range of ability levels were randomly played to raters. The first video portrayed a counselor with fundamental counseling skills and a number of potentially damaging behaviors. The second video depicted a counselor with superior counseling abilities and no possibly damaging behaviors. The third counselor possessed more fundamental counseling abilities and fewer potentially damaging behaviors. The final film depicted a psychotherapist with no basic counseling skills and potentially detrimental client interactions.

The videos were pre-recorded sessions of a professional counselor and an actor following a script that addressed the competency levels listed above. Raters scored each video independently and submitted the ENACT tool to the trainer for inter-rater reliability calculation. The trainer requested that each rater provide their scores and justifications. After discussing the scores, inconsistencies were resolved. The instructor provided the expert scores, rationale, and assured that the raters comprehended the counselor's actions and the correct score to offer them.

Virtual training for the Common Elements Treatment Approach in telehealth (T-CETA)

In September 2020, a one-day virtual rater training for a telehealth-adapted ENACT version was performed by telephone. Prior to the training, the raters were provided with training materials (ENACT measure, headphones, paper, and pen). The trainer launched a conference call and advised participants to sit in a quiet, distraction-free environment. The training began with a thorough review of the telehealth-adapted ENACT tool to verify that raters remembered each item from the previous training and understood the adjustments, if relevant.

The trainer then selected two raters to play a counselor and a client for four to six minutes while the remaining three raters listened and scored using the ENACT. After each role-play, the trainer selected two additional raters to ensure that each rater engaged in listening to and scoring the role-plays. The trainer recited the same acting cues used during the in-person instruction. The trainer requested that each rater provide their scores and justifications for each role-play. The instructor noted differences between each rater's results and the expert's scores. The instructor permitted the raters to discuss their scores prior to revealing the expert scores and explanation.

Similar to the in-person instruction, the trainer created role-plays based on counselor behaviors that raters have difficulty accurately and consistently scoring. One of the raters played the actor, while the trainer portrayed a counselor. The other raters listened to the role-plays and assigned scores. The instructor requested that each rater share their results with the group prior to providing the group with the expert scores and reasons.

Counseling sessions captured on tape

The raters listened to and scored the initial audio recordings of counseling sessions before texting their scores to the trainer for inter-rater reliability computation. In each of the four audio files, a counselor

introduced counseling to a client. The audio recordings were played in a random order for the raters. With the exception of one item, the counselor in the first audio had advanced counseling abilities on the majority of things. In the second recording, the counselor demonstrated an equal mix of advanced and fundamental counseling abilities on the majority of things, with only one issue being insufficiently described. The third audio was a recording of a counselor demonstrating all the basic counseling skills on some items and a few on others. The fourth audio recording captured a counselor with the greatest levels of potentially dangerous behaviors on the majority of goods, as well as a mix of all the basic abilities on certain products and advanced skills on a few items.

After listening to an audio recording, raters had five minutes to provide a score. The instructor terminated the conference call so that raters could text their scores. The trainer then launched the conference call after calculating their inter-rater reliability. The instructor requested that each rater provide their scores and justifications. The raters discussed each other's scores before the trainer provided the expert scores and justifications.

Data analysis

Intra-class correlations (ICC) estimates were calculated using the Mangold Reliability Calculator (Program Version 1.5, Lab Suite Version 2015) based on a mean-rating (k=5), 2-way random-effects model for all four videos from the in-person training and all four audio recordings from the virtual training. Inter-rater reliability of rater scores was assessed using inter-class correlation coefficients (ICC's). Koo and Li[22] categorize ICC scores as follows: below 0.5 is poor consistency; between 0.5 and 0.75 is moderate consistency; between 0.75 and 0.9 is a good consistency and above 0.9 is excellent.

Data from the in-person training were collected and analysed during the in-person training. Similarly, data from the telephone training were collected and analysed during the telephone training. Each of the 5 raters using the 15-item ENACT measure scored the same role-plays that they watched and listened to during the two trainings. Table 2 summarizes the process of rater training.

RESULTS

ICC scores for the in-person training (ranging from 0.81 to 0.89) showed good consistency and the scores for the telephone training (ranging from 0.71 to 0.85) showed moderate consistency for one audio and good consistency for two audios. Negative ICC scores were computed for the second video and audio. The negative ICC scores were caused by a lack of variation in the skill levels for different competencies in the role-play (i.e., the counselor in the video performed all skills at the same level of competency). For the second video, the raters scored 9 items out of the 15 exactly the same (i.e., perfect agreement with ICC=1). For the second audio, the raters scored four items exactly the same (i.e., perfect agreement) and four items almost exactly the same with only one rater giving a different score from the other four raters. See Table 3 for ICC scores.

DISCUSSION

This study describes a brief training strategy for mental health professionals with no prior experience using rating scales to assess the competency of lay counselors using the ENACT measure. Findings indicate that the training approach was successful in achieving high inter-rater reliability, with ICC ranging from moderate to good (ICC: 0.71-0.89). Training for raters reduces error and variability in clinical outcome assessments.[23] Inter-rater reliability is contingent on rater consistency and modified by rater selection and training.[24] There are no agreed-upon selection and training criteria for raters.[25] In this study, raters were chosen based on their prior CETA training and their willingness to get more training. Rohan et al.[26] advocate the use of competent raters for the clinical intervention being scored. When raters evaluate outcomes outside their area of expertise, both reliability and validity are impaired.[27] In this study, raters had previously administered CETA for an average of three to four years and were

familiar with the constructs they were scoring.

The in-person and telephone training sessions commenced with a review of the ENACT instrument. This review was instructive and featured item definition, item scale, and item differentiation. Item evaluation is acknowledged as a crucial phase in rater training.[7] Item review aids in addressing misinterpretation of conceptions, terminology, and the scale, which can lead to excessive score variation among raters.[22] The instructor instructed the raters on scoring conventions and verified their comprehension of the ENACT items. The trainer encouraged the raters to ask questions and provide feedback. The trainer took note of individual rater questions, reviewed them with all trainees, and included them into their role-plays. This is consistent with other rater training models that document individual rater questions and incorporate them into subsequent role-play scenarios.[13]

The scoring of standardized role-plays with trainer observation and feedback was a significant component of both in-person and telephone trainings. The raters utilized the ENACT instrument to assess the ability of their fellow trainees during a standard role-play based on a scenario provided by the instructor. The instructor observed role-plays, assessed them, compared rater and expert results, and then discussed accurate and outlier scores with each group. This method is consistent with existing research on training models, including recommendations that a trainer co-rate with the raters until they are competent and the use of an interactive round table teaching technique as opposed to a classroom-style approach.[28,25]During this portion of training, raters were encouraged to converse with one another, and the trainer took note of difficult-to-score counselor behaviors and remarks. During training, the trainer improvised additional role-plays based on difficult-to-score counselor behaviors so that raters may gain more experience scoring these behaviors.

Due to inadequate internet access, a telephone conference call was preferred over a video call for the September 2020 training. Before the start of research studies, it is common for remote role-playing exercises to be done.[27] This sort of practical training allowed the trainer to analyze the raters and determine which ENACT issues were the most challenging to evaluate and required more didactic teaching and role-play scenarios. Applied learning is a well-established method that has been utilized in a variety of sectors to train raters.[8,26]

Scoring the video and audio recordings was the final step of the in-person and telephone training. After calculating ICC scores, the trainer conducted a simulated interview with each rater to evaluate their accuracy and justification. In addition, as part of the training, Rohan et al.[26] evaluated raters using simulated interviews. When the trainer offered the expert scores and rationale, the video and audio recordings could be replayed so that specific counselor behaviors could be referenced. Asan and Montague[29] propose using films since they may be replayed and provide additional information that cannot be obtained through in-person observations. It is well acknowledged that video and audio recordings are scored by raters.[11,30,31] This research has taught us that standardized role-plays must incorporate a range of skill levels across skills (e.g. some competencies performed at level 1, some at level 2). If a conventional role-play comprises all skills at the same level (for example, all skills at level 3), this does not introduce enough variance to calculate ICC values accurately.

Limitations

This study was limited by the small number of raters that participated. To train a larger number of raters using this approach will require more trainers in order to maintain adequate trainer observation and feedback given to each individual rater. Additionally, the lack of variability in displayed skill levels in some videos and role-plays resulted in the computation of negative ICC scores that are difficult to interpret. It has been argued that the use of video and audio recordings can artificially raise ICC scores because raters view the video or audio recordings at the same time (i.e., with no information variance) when in practice they may score different people at different times.[27,31] Finally, ICC data for scoring conducted after the rater trainings were not collected, which limits our ability to report maintenance of rating skills over time. Recommendations for future

studies include the application of this training model with a larger sample of raters and collecting longitudinal rater consistency scores to test maintenance of skills over time which can inform whether additional "booster" trainings are needed to maintain high consistency. Future studies should also collect "real world" competency scoring of counselors in sessions with clients to determine if raters can maintain high consistency in live sessions.

CONCLUSIONS

The training model outlined here was effective in achieving high consistency of counselor competency scores for lay providers with no previous rating experience. Findings suggest lay providers can be trained to use the ENACT measure to score counselor competency for clinical research and program monitoring. Effective training of competency raters should include: 1) didactic learning through complete item review and instruction on scoring conventions, 2) practical learning with trainer observation, co-rating, elicitation of scoring rationale and feedback, and 3) the use of video or audio recordings, which contain variation in skill levels, to assess rater reliability. This training model can be adapted for use in other studies and programs with different rating scales. Training raters to assess counselor competency with rating scales such as the ENACT tool is essential to ensure evidence-based mental health treatments are administered with quality and fidelity.

INFORMATION

Acknowledgements. The authors wish to thank the raters and actors for all their hard work and willingness to learn new skills throughout the EQUIP study. We further wish to thank and acknowledge the counselors that agreed to participate in multiple standardized role-plays, which were critical for this research. We are grateful to the individuals who participated in the studies in each of the EQUIP sites and all members of the EQUIP team for their dedication, hard work and insights.

Authors' contributions. MMM: Writing initial draft, data analysis, review and editing, project manager; CF: Writing initial draft, project design, data analysis, manuscript review and editing; KM: Project design, project oversight, manuscript review and editing, clinical supervision, co-investigator; JCK: Project design, manuscript review and editing, data analysis; SSVW: Project design, manuscript review and editing, clinical supervision; SMM: Manuscript review and editing, project management; IS: Project design, manuscript review and editing, principle investigator; LM: Project design, manuscript review and editing, principle investigator; GP: Project design, manuscript review and editing, data analysis, EQUIP investigator; BK: Project design, manuscript review and editing, data analysis, EQUIP investigator, final review before submission.

Conflict of interest. The authors declare no potential conflict of interest.

Funding. Funding for the WHO EQUIP initiative is provided by USAID. JCK's contribution was supported in part by a grant from National Institute on Alcohol and Alcoholism (NIAA; K01AA026523). The authors alone are responsible for the views expressed in this article and they do not represent the views, decisions or policies of the institutions with which they are affiliated.

Availability of data and materials. All data generated or analyzed during this study are included in this published article.

Ethical Considerations. Ethical approval for this study was granted by the University of Zambia Biomedical and Research Ethics Committee (UNZABREC) approval number – ref-008-02-19, Johns Hopkins School of Public Health (JHSPH; IRB No. 00009259), George Washington University (GWU; IRB No. NCR191797), and the World Health Organization (WHO; ERC.0003192). Lay mental health providers interested in being trained as competency raters provided written consent to participate in the study. All data collected from raters were anonymized and all hard copy forms stored in a locked cabinet.

REFERENCES

1. Santomauro DF, Herrera AM, Shadid J, et al. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet. 2020; 398. https://doi.org/10.1016/S0140-6736(21)02143-7

2. Rhem J, Shield KD. Global Burden of Disease and the Impact of Mental and Addictive Disorders. Curr. Psychiatry Rep. 2019; 21(10). https://doi.org/10.1007/s11920-019-0997-0

3. Wainberg ML, Scorza P, Shultz JM, et al. Challenges and Opportunities in Global Mental Health: a Research-to-Practice Perspective. Curr Psychiatry Rep. 2017; 19(5): 28-34. doi: https://doi.org/10.1007/s11920-017-0780-z

4. Lund C, Alem A, Schneider M, et al. Generating evidence to narrow the treatment gap for mental health disorders in sub-saharan Africa: rationale, overview and methods of AFFIRM. Epidemiol Psychiatri Sci. 2015; 24: 233-240. doi: https://doi.org/10.1017/S2045796015000281

5. Branson A, Myles P, Mahdi M, Shafran R. The relationship between competence and patient outcome with low intensity cognitive behavioural interventions. Behav cogn psychother. 2018; 46(1): 101-114. https://doi.org/10.1017/S1352465817000522

6. Liness S, Beale S, Lea S, et al. Multi professional IAPT CBT Training: Clinical Competence and Patient Outcomes. Behav and Cogn Psychother. 2019. https://doi.org/10.1017/S1352465819000201

7. Brown RC, Southern-Gerow M, McLeod BD, et al. The Global Therapist Competence Scale for Youth Psychosocial Treatment: Development and Initial Validation. J. Clin. Psychol. 2019; 74(4): 649-664. Doi: https://doi.org/10.1002/jclp.22537

8. McLeod BD, Southam-Gerow MA, Rodriguez A, et al. Development and Initial Psychometrics for a Therapist Competence Instrument for CBT for Youth Anxiety. J Clin Child Adolesc Psychol. 2018; 47(1): 47-60. doi: https://doi.org/10.1080/15374416.2016.1253018

9. Muse K, McManus F. A systematic review of methods for assessing competence in cognitive-behavioural therapy. Clin. Psychol. Rev. 2013; 33(3): 484-499. doi: https://doi.org/10.1016/j.cpr.2013.01.010

10. Bjaastad JF, Haugland BSM, Fjermestad KW, et al. Competence and Adherence Scale for Cognitive Behavioral Therapy (CAS-CBT) for Anxiety Disorders in Youth: Psychometric Properties. Psychological Assessment, *28*(8), 908–916. https://doi.org/10.1037/pas0000230

11. Ottman K, Kohrt BA, Pedersen G, Schafer A. Use of Role-plays to Assess Therapist Competency and its Association with Client Outcomes in Psychological Interventions: A Scoping Review and Competency Research Agenda. 2019. https://www.sciencedirect.com/science/article/pii/SOOO5796719302177

12. Kohrt BA, Jordans MJD, Rai S, et al. Therapist competence in global mental health: Development of the Enhancing Assessment of Common Therapeutic factors (ENACT) rating scale. Behav Res and Ther. 2015; 69: 11-21. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4686771

13. Glanzman AM, Mazzone ES, Young SD, et al. Evaluator Training and Reliability for SMA Global Nusinersen Trials. J. Neuromuscul Dis. 2018; 5: 159-166. doi: https://doi.org/10.3233/JND-180301

14. Schmidt LK, Andersen K, Nielsen AS, Moyers TB. Lessons learned from measuring fidelity with the Motivational Interviewing Treatment Integrity code (MITI 4). J. Subst. Abuse Treat. 2019; 97: 59-67

15. Kohrt BA, Schafer A, Willhoite A, et al. Ensuring Quality in Psychological Support (WHO EQUIP): developing a competent global workforce. World Psychiatry. 2020; 19:115-6. doi: https://doi.org/10.1002/wps.20704

16. Murray LK, Dorsey S, Haroz E, et al. A common elements treatment approach for adult mental health problems in low- and middle-income countries. Cogn. Behav. Pract. 2014; 21:111-123. https://doi.org/10.1016/j.cbpra.2013.06.005

17. Murray LK, Kane JC, Glass N, et al. Effectiveness of the Common Elements Treatment Approach (CETA) in reducing intimate partner violence and hazardous alcohol use in Zambia (VATU): A randomized controlled trial. PLoS Med. 2020; 17(4). https://doi.org/10.1371/journal.pmed.1003056

18. United Nations, Department of Economic and Social Affairs, Population Division. World Population Prospects. 2019. Volume 2: Demographic Profiles (ST/ESA/SER.A/427).

19. Munakampe NM. Strengthening Mental Health Systems in Zambia. Int. J. Ment. Health Syst. 2020; 14(28). https://doi.org/10.1186/s13033-020-00360-z

20. Kohrt BA, Mutamba BB, Luitel NP, et al. How competent are non-specialists trained to integrate mental health services in primary care? Global perspectives from Uganda, Liberia, and Nepal. Int. Rev. Psychiatry. 2018; 30(6): 182-98. https://doi.org/10.1080/09540261.2019.1566116

21. Kohrt BA, Ramaiya MK, Rai S, et al. 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. GMH. 2015; 2(23): 1-16. https://doi.org/10.1017/gmh.2015.21

22. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016 Jun;15(2):155-63. doi: 10.1016/j.jcm.2016.02.012. Epub 2016 Mar 31. Erratum in: J Chiropr Med. 2017 Dec;16(4):346.

23. Sadler ME, Yamamoto RT, Khurana L, Dallabrida SM. the impact of rater training on clinical outcomes assessment data: a literature review. Int. J. Clin. Trials. 2017; 4(3): 101-110. https://www.sprim.com/wp-content/uploads/2020/06/Sadler-2017-Rater-training-rvw.pdf

24. Shweta, Bajpai RC, Chaturvedi HK. Evaluation of Inter-Rater Agreement and Inter-Rater Reliability for Observational Data: An Overview of Concepts and Methods. J. Indian Acad. Appl. Psychol. 2015; 41(3) Special issue: 20-27.

25. West MD, Daniel DG, Opler M, et al. Consensus Recommendations on Rater Training and Certification. Innov. Clin. Neurosci. 2014; 11(11): 10-13.

26. Rohan KJ, Rough JN, Evans M, et al. A protocol for the Hamilton Rating Scale for Depression: Item Scoring Rules, Rater Training, and Outcome Accuracy with Data on its Application in a Clinical Trial. J. Affect. Disord. 2016; 200: 111-118. doi: https://doi.org/10.1016/j.jad.2016.01.051

27. Kobak K, Williams JBW, Engelhardt N, Lipsitz J. Rater Training in Multicenter Clinical Trials. J. Clin. Psychopharmacol. 2004.Available from: doi: https://doi.org/10.1097/01.jcp.0000116651.91923.54

28. Opler MGA, Yavorsky C, Daniel DG. Positive and Negative Syndrome Scale (PANSS) Training: Challenges, Solutions, and Future Directions. Innov Clin Neurosci. 2017; 14(11): 77-81.

29. Asan O, Montague E. Using video-based observation research methods in primary care health encounters to evaluate complex interactions. Informatics in primary care. 2014; 21: 161-170. Doi: https://doi.org/10.14236/jhi.v21i4.72.

30. Gulgin H, Hoogenboom B. The Functional Movement Screening (FMS): An inter-rater reliability study between raters of varied experience. Int. J. Sports Phys. Ther. 2014; 9(1): 1-7

31. Sajatovic M, Gaur R, Tatsuoka C, et al. Rater training for a multi-site, international clinical trial: What mood symptoms may be most difficult to rate? Psycho Pharmacol Bull. 2011; 44(3):5-14.

How to cite this article: Mwenge M.M., Figge C.J., Metz K., Kane J.C., Kohrt B.A., Pedersen G.A., Sikazwe I., Skavenski Van Wyk S., Mulemba S.M., Murray L.K. **Improving inter-rater reliability of the enhancing assessment of common therapeutic factors (ENACT) measure through training of raters**. Journal of Public Health in Africa. 2022;13:2201. https://doi.org/ 10.4081/jphia.2022.2201

TABLE 1: Rater socio-demographic characteristics.

	n	%
Age (Median, Range)	36	28 - 66
Gender		
Male	1	20
Female	4	80
Education		
School certificate	2	40
College certificate	1	20
University Diploma	2	40
Counseling Experience prior to study		
2 years and below	1	20
3 - 4 years	3	60
5 years and above	1	20
Rater experience prior to study		
No experience	5	100

TABLE 2: Training method.

Training Procedure	Activity Descriptions
Didactic learning	Overview of study background ENACT item review i.e. item definitions, review of item rating scales, differentiation of rating scales
Practical learning	Group role-plays Trainer observation and scoring Comparison of expert scores to trainee scores Trainer rationale and reconciliation of score Discussion of scoring conventions
Trainer group feedback and role-play	Group feedback on inaccurate scores Trainer role-play with trainee queries incorporated Comparison of expert scores to trainee scores Trainer rationale and reconciliation of scores
Video/audio reliability scoring	Raters watch/ listen to video or audio independently Trainer calculates ICC scores for raters Mock interview for raters on their scores and rationale Trainer rationale and reconciliation of scores

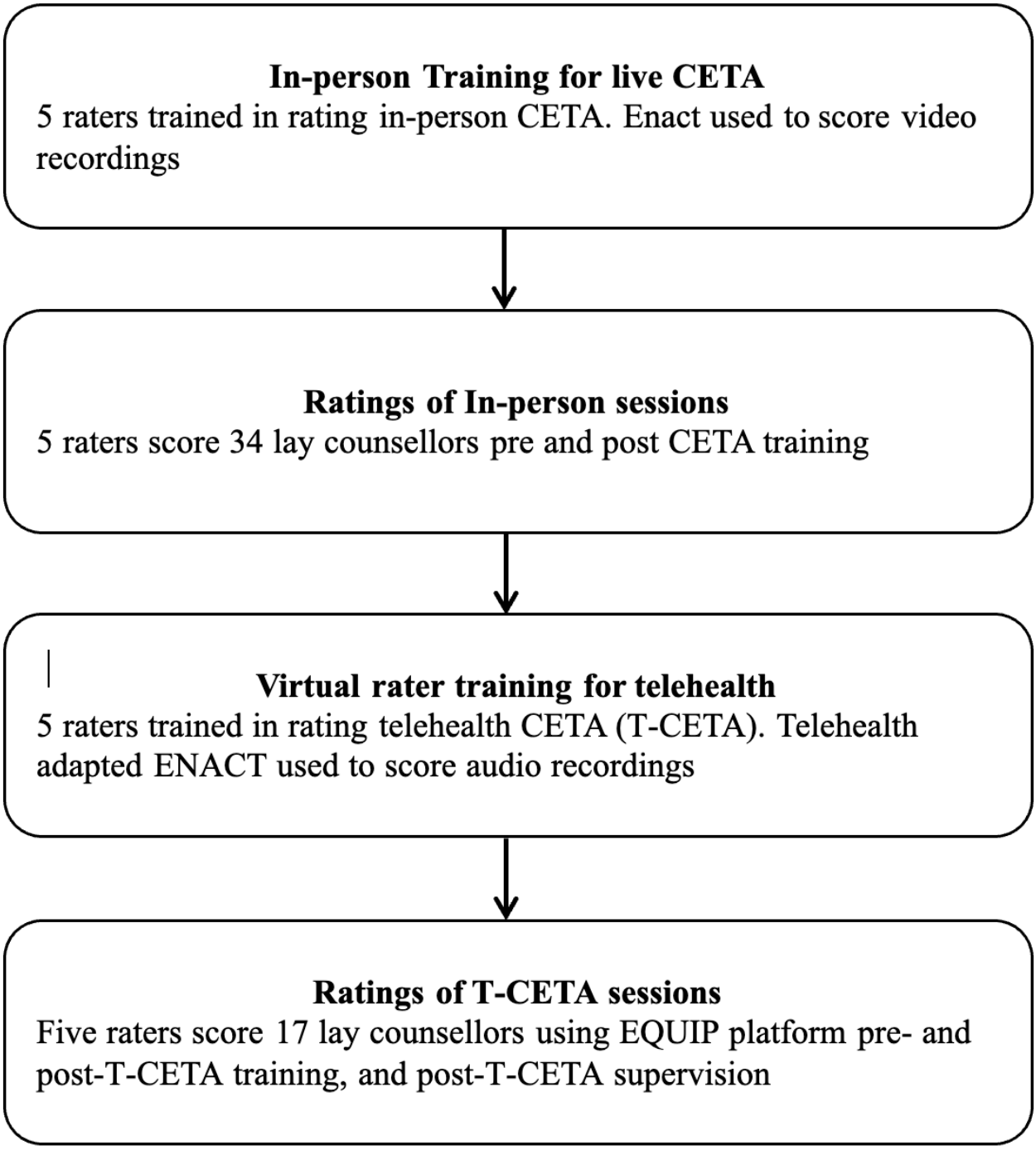


FIGURE 1: Study procedures flow diagram.

TABLE 3: ICC Scores for in-person and telephone trainings.

In-person Training	ICC Score
Video 1	0.89
Video 2	-1.74
Video 3	0.81
Video 4	0.82
Telephone Training	
Audio 1	0.71
Audio 2	-1.35
Audio 3	0.81
Audio 4	0.85