global mental health

TEACHING AND LEARNING

ORIGINAL RESEARCH PAPER

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

B. A. Kohrt[1–3]*, M. K. Ramaiya[3,4], S. Rai[1,3], A. Bhardwaj[3] and M. J. D Jordans[5,6]

[1] *Transcultural Psychosocial Organization (TPO) Nepal, Kathmandu, Nepal*
[2] *Department of Psychiatry and Behavioral Sciences, Duke University, Durham, USA*
[3] *Duke Global Health Institute, Duke University, Durham, USA*
[4] *Department of Psychology, University of Nevada, Reno, USA*
[5] *HealthNetTPO, Amsterdam, the Netherlands*
[6] *Centre for Global Mental Health, Institute of Psychiatry, Psychology, and Neuroscience, King's College London, London, UK*

Global Mental Health (2015), **2**, e23, page 1 of 16. doi:10.1017/gmh.2015.21

Background. Task-sharing is the involvement of non-specialist providers to deliver mental health services. A challenge for task-sharing programs is to achieve and maintain clinical competence of non-specialists, including primary care workers, paraprofessionals, and lay providers. We developed a tool for non-specialist peer ratings of common factors clinical competency to evaluate and optimize competence during training and supervision in global mental health task-sharing initiatives.

Methods. The 18-item **EN**hancing **A**ssessment of **C**ommon **T**herapeutic factors (ENACT) tool was pilot-tested with non-specialists participating in mental health Gap Action Programme trainings in Nepal. Qualitative process evaluation was used to document development of the peer rating scoring system. Qualitative data included interviews with trainers and raters as well as transcripts of pre- and post-training observed structured clinical evaluations.

Results. Five challenges for non-specialist peer ratings were identified through the process evaluation: (1) balance of training and supervision objectives with research objectives; (2) burden for peer raters due to number of scale items, number of response options, and use of behavioral counts; (3) capturing hierarchy of clinical skills; (4) objective *v.* subjective aspects of rating; and (5) social desirability when rating peers.

Conclusion. The process culminated in five recommendations based on the key findings for the development of scales to be used by non-specialists for peer ratings in low-resource settings. Further research is needed to determine the ability of ENACT to capture the relationship of clinical competence with client outcomes and to explore the relevance of these recommendations for non-specialist peer ratings in high-resource settings.

Received 23 February 2015; Revised 16 October 2015; Accepted 25 October 2015

Key words: Competence, culture, global health, measurement, task sharing, training.

* Address for correspondence: B. A. Kohrt, MD, PhD., Duke Global Health Institute, 213 Trent Hall, 310 Trent Drive, Durham, NC 27708, USA.
(Email: brandon.kohrt@duke.edu)

https://doi.org/10.1017/gmh.2015.21 Published online by Cambridge University Press

Introduction

Task-sharing is the involvement of non-specialist providers in the delivery of mental health services that are, at present, predominantly performed by mental health professionals (Patel, 2009, 2012; WHO, 2008). In global mental health, non-specialist providers represent a broad category of professional and lay cadres who lack formal clinical training in fields such as psychology, psychiatry, and social work. Non-specialists may include social and community health workers, peer volunteers, teachers, midwives, traditional healers, and others lacking professional training experience. In low-and-middle-income countries (LMICs), examples of task-sharing include training community health workers and persons without a professional health role to diagnose and manage common mental disorders and deliver specialized psychological treatments including interpersonal therapy, behavioral activation, and cognitive processing therapy (van Ginneken *et al.* 2013).

Lack of properly trained personnel is a key barrier to dissemination, implementation, and scaling up of task-sharing programs (Murray *et al.* 2014*a*). The limited number of mental health specialists presents a challenge not only for evidence-based delivery of mental health care but also for training and supervision. Clinical resources for training and supervision—both in terms of human capacity and materials—are historically lacking in most LMIC settings (Abas *et al.* 2003; Kakuma *et al.* 2011). Therefore, there is a need for task-share quality improvement activities with non-specialist providers as well.

One means of addressing this barrier to scaling up services in global mental health is utilizing culturally adapted tools to evaluate therapist competence. Therapist competence reflects how well a health worker implements a technique and implies both adhering to an evidence-based treatment and implementing it with appropriate skill and is typically measured with structured role plays using standardize patients; therapy quality refers to how well a treatment is delivered in a clinical setting with real patients (Margison *et al.* 2000; Fairburn & Cooper, 2011). Without easily administered assessments of non-specialist competence, it is challenging to determine both the minimum skill level necessary for effective delivery of evidence-based treatments, as well as the success of training and supervision practices on the development and maintenance of these competencies. One innovative approach is the use of instruments to assess therapist competence that can be administered by non-specialist peers to promote quality improvement (Singla *et al.* 2014), rather than using tools requiring trained experts for administration and scoring.

In high-income settings, there are a wide variety of instruments available for expert ratings of therapeutic competence and quality (Margison *et al.* 2000; Cahill *et al.* 2008). These vary in length, with some tools containing 12 items, e.g. Cognitive Therapy Scale-Revised, CTS-R (Blackburn *et al.* 2001), and others containing up to 70, e.g. Hill Interaction Matrix-Group, HIM-G (Hill & Gormally, 1977), or more, e.g. Vanderbilt Therapeutic Alliance Scale (O'Malley *et al.* 1983), (see Table 1). Response options are typically Likert scales that reflect the degree of agreement (e.g. 'strongly disagree' to 'strongly agree') with a declarative statement. Other tools include frequency markers or behavior counts. Response alternatives vary from five-point scales, e.g. Cross-Cultural Counseling Inventory-Revised, CCCI-R (Lafromboise *et al.* 1991), to nine-point scales, e.g. Therapist post-session Questionnaire (Samstag *et al.* 1998), with most containing seven response options.

Our goal was to develop a novel rating scale that would be appropriate for peer ratings by non-specialists in a global mental health context. Despite the abundant existence of peer rating scales in high-income settings, there are a number of concerns that these instruments may have limited applicability and relevance in LMICs. For instance, Likert scale peer rating instruments capture subjective feelings relative to a non-representative comparison group, thereby limiting their ability to capture absolute or context-free judgment at the population level (Biernat *et al.* 1991). Research in cross-cultural psychology echoes these concerns regarding Likert scales (Heine *et al.* 2002). Further, instruments containing too many response alternatives are associated with increased cognitive load, fatigue, and response bias (Anastasi, 1976; Krosnick, 1991), as well a decrease in ability to accurately discriminate between options (Miller, 1956). These concerns may be magnified for non-specialists with limited-to-no prior training in psychological instrument administration.

Dissemination and implementation theory calls for tools to be context-appropriate (Murray *et al.* 2014*a*), thus requiring task-sharing instruments in LMICs to be consistent with the objectives and resources of the prevailing cultural group. Across studies and interventions 'common factors' refer to those processes, which contribute to positive patient outcomes regardless of specific treatment features. Although definitions vary (Frank & Frank, 1991; Wampold, 2011), common factors often refer to therapist qualities and therapist–client interactions including empathy and genuineness, client and extra-contextual factors such as mobilization of social support and bolstering of prior positive coping strategies, and aspects of the therapist–client relationship including collaborative goal setting and

https://doi.org/10.1017/gmh.2015.21 Published online by Cambridge University Press

Table 1. *Examples of response options for therapist rating scales*

Tool	Number of items	Response options
Coding the Interaction in Psychotherapy, CIP (Schindler, Hohenberger-Sieber, & Hahlweg, 1989)	19 items	Relative frequency response with behavior counts
CTS-R (Blackburn *et al.* 2001)	12 items	0–6 scale, with skill-based descriptions for each response presented in hierarchy of mastery
CRF-S (Corrigan & Schmidt, 1983)	12 items	7-point scale with blank spaces for 'X' marking, with one anchor point at either end for 'not very' and 'very'
CCCI-R (Lafromboise *et al.* 1991)	18 items	5-point Likert scale, 'strongly disagree' to 'strongly agree'
Helping Alliance Questionnaire revised, HAq-II (Luborsky *et al.* 1996)	19 items	1–6 Likert scale, 'strongly disagree' to 'strongly agree' with anchor points for each score
HIM-G (Hill & Gormally, 1977)	72 items	0–5 scale, with various response sets; e.g. frequency 'not at all', '0–10% of the time', … 'over 60% of the time'; number of people 'most people,' 'few people', … 'nobody'
Jefferson Empathy Scale, JES (Suh, Hong, Lee, Gonnella, & Hojat, 2012)	20 items	7-point Likert scale, 'strongly disagree' to 'strongly agree'
Maslach Burnout Inventory, MBI (Maslach, Jackson, & Leiter, 1996)	22 items	0–6 scale for frequency, '0 = never', '6 = everyday'
Rater Applied Performance Scale, RAPS (Lipsitz *et al.* 2004)	6 items	5-point scale, 'not applicable', 'unsatisfactory', 'fair', 'good', 'excellent'
Therapist Action Tool, TAT (Hoyt, Marmar, Horowitz, & Alvarez, 1981)	25 items	0–5 scale with four anchor points, '0 = did not do it', '1 = occurred but minor', '3 = moderate', '5 = major emphasis'
Therapist Post-Session Questionnaire (Samstag *et al.* 1998)	40 items	A variety of response sets, e.g. 1–9 Likert scale, categorical responses, open-ended, and 7-point dyadic attributes (bad to good, safe to dangerous)
Vanderbilt Therapeutic Strategies Scale, VTSS (Henry *et al.* 1993)	21 items	5-point scale with frequency responses for some items and degree of quality responses for other items

overall promotion of hope and expectancy of change among clients (Garfield, 1973; Greenberg, 2004; Sparks *et al.* 2008; Karson & Fox, 2010; Wampold, 2011).

Despite this importance of common factors for positive patient outcomes, measurement of therapist competence related to common factors in LMICs has been limited, with few researchers systematically adapting or developing novel tools for use by non-specialists (Kabura *et al.* 2005). Working with non-specialists in Uganda, Kabura *et al.* (2005) adapted the Attending Behavior Rating Scale (Ivey & Authier, 1978), which is used to rate micro-counseling skills: eye contact, vocal tone, posture, and verbal attending behavior. The Uganda adaptations included adding specific frequency counts for behaviors such as open-ended questions, paraphrasing, and reflecting feelings. However, no cultural adaptations were described. Studies using World Health Organization (WHO) recommendations for mental health integration in primary care have included observed evaluations of therapy quality in actual clinical practice, but role plays with standardized patients for competence evaluation were not reported (Sadik *et al.* 2011; Makanjuola *et al.* 2012; Jenkins *et al.* 2013). Overall, there has been a lack of adaptations and novel tool development for non-specialists to evaluate competence in cross-cultural settings.

To address this gap in availability of tools for global mental health, we have developed a tool to use during training and supervision in LMICs to assess common factors competence among non-specialist providers delivering mental health services (Kohrt *et al.* 2015). This tool, the **EN**hancing **A**ssessment of **C**ommon **T**herapeutic factors (ENACT) rating scale, was developed using a four-step systematic process in Nepal. We have previously described the systematic process of item generation, piloting, and basic psychometric properties (Kohrt *et al.* 2015).

Objective

In this paper, we focus on the development of the scoring system for ENACT, with the goal of describing the process to optimize feasibility and utility among non-specialist peer raters in a LMIC cultural context. We

https://doi.org/10.1017/gmh.2015.21 Published online by Cambridge University Press

present a qualitative process evaluation that led to a three-tiered competency rating system designed for use in task-sharing initiatives in global mental health. The findings are relevant to guide development of non-specialist peer ratings for global mental health in low-resource settings.

Methods

Tool development

ENACT was developed in the context of the Program to Improve Mental Health Care (PRIME), an initiative in LMICs to develop mental health services in primary care and community settings (Lund *et al.* 2012; Jordans *et al.* 2013). In Nepal's Chitwan District, primary care and community health workers are being trained with a locally developed Mental Health Care Package (Jordans *et al.* 2015), which includes WHO's mental health Gap Action Programme (*mhGAP*) *Intervention Guide* (*mhGAP-IG*) (WHO, 2010), psychosocial skills modules, and brief modified versions of behavioral activation (the Healthy Activity Program) and motivational interviewing (Counseling for Alcohol Program) (Patel *et al.* 2014; Singla *et al.* 2014). The target settings in Nepal include health posts, sub-health posts, and primary health centers. These are the first port of entry for the general public seeking medical care. All trainees in PRIME are non-specialists working in these settings. The trainers for primary care workers include Nepali psychiatrists trained on mhGAP-IG and Nepali psychosocial counselors with a decade of experience in mental health programs and more than 5 years of clinical supervision work.

ENACT (see full tool in online Supplemental Material) was developed using a four-step process within PRIME (Kohrt *et al.* 2015). First, a review of 56 client–therapist interaction instruments was used to generate a pool of common factors. Role-plays between Nepali therapists and standardized clients were used to generate additional domains of cultural and clinical significance. Then 10 Nepali therapists scored the initial domains for comprehensibility and clinical significance. Domains with high scores in both of comprehensibility and clinical significance were pilot tested with two expert Nepali therapists using brief standardized role-plays. Session transcripts were then qualitatively coded to examine the instrument's feasibility, acceptability, and reliability. After further domain consolidation and revision, a list of 18 unique items was generated. The final 18-item tool incorporated a 3-tiered response system 1 ('needs improvement'), 2 ('done partially'), and 3 ('done well'). Psychometric measurements of the 18-item instrument followed. The intra-class correlation coefficient (ICC) for expert therapists was established through rating videotaped sessions: ICC (2,7) = 0.88 (95% CI 0.81–0.93). Non-specialist peer ICC (1,3) based on post-training role-plays was 0.67 (95% CI 0.60–0.73). Cronbach's alpha based on 34 expert ratings of non-specialist roles plays was 0.89. Cronbach's alpha for non-specialist peer-ratings was 0.80 (*N* = 113).

Process evaluation

In this paper, we present the qualitative process evaluation that led to development of a three-tiered response structure designed to optimize feasibility and utility for peer raters. We describe the challenges encountered during scoring system development and how these were addressed. Scoring guidelines were created in an iterative process conducted in tandem with the overall ENACT protocol (See Figure 1).

Development of an initial scoring system

After Step 1 (domain generation) and Step 2 (domain relevance) of ENACT development, a 49-item tool was created for piloting. A three-tiered scoring system containing the values 0 'not at all', 1 'minimal use', and 2 'effective use' was selected for rating (see Figure 2). We initially chose a three-point scale as the minimally sufficient response option set. If the three-point scale had high feasibility and utility in proceeding steps, we would conserve it. If, however, it were deemed inadequate, additional points could be added to achieve the minimum number of response options required to capture variance and change in competence throughout the training and supervision process. This strategy was chosen based on prior literature suggesting cognitive burden and discriminatory impairments

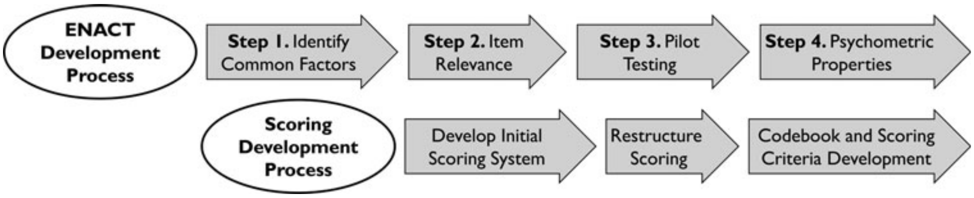
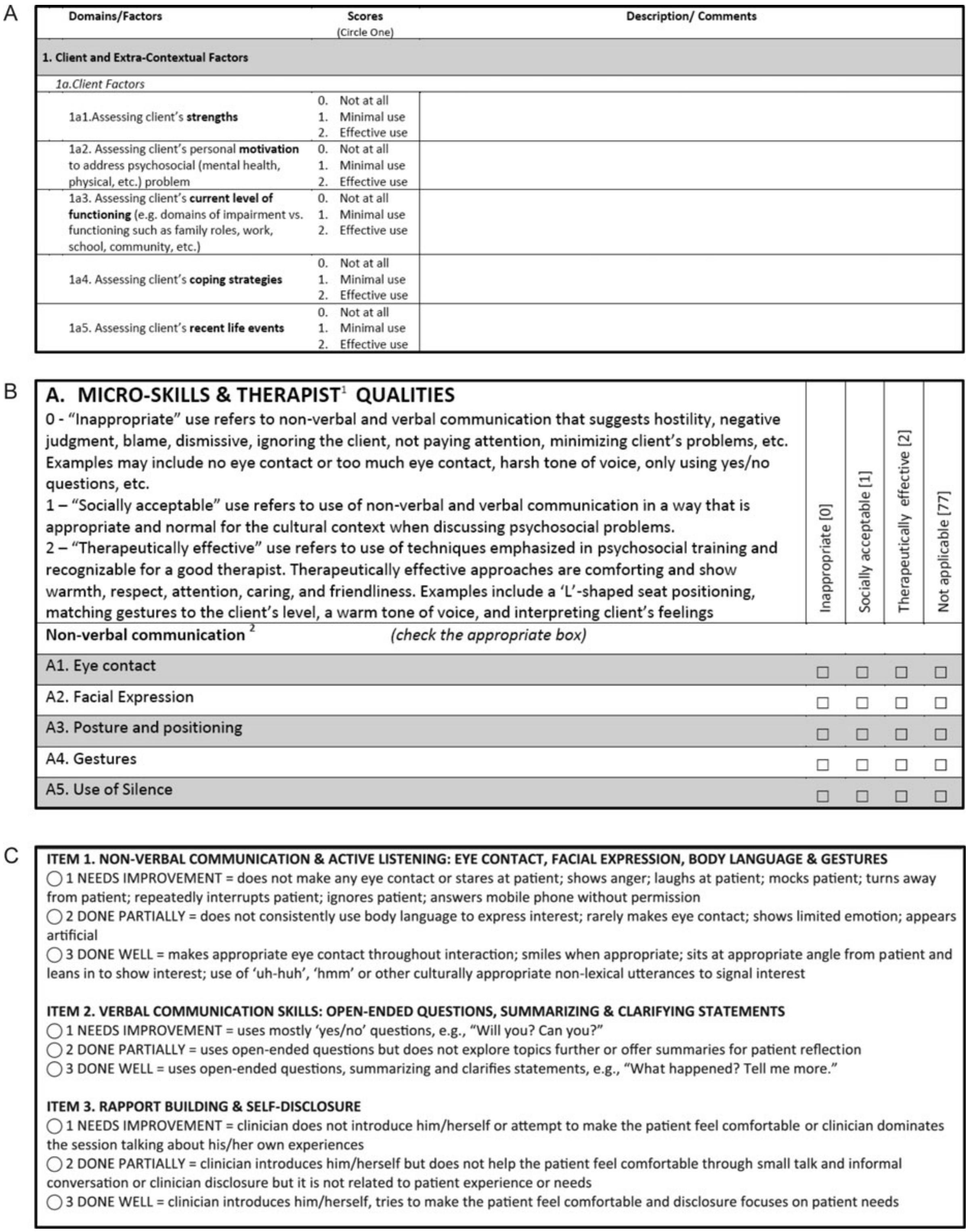

ENACT Development Process
Step 1. Identify Common Factors
Step 2. Item Relevance
Step 3. Pilot Testing
Step 4. Psychometric Properties
Scoring Development Process
Develop Initial Scoring System
Restructure Scoring
Codebook and Scoring Criteria Development

Fig. 1. Scoring guideline development process within the ENACT tool development process.

https://doi.org/10.1017/gmh.2015.21 Published online by Cambridge University Press



A

Domains/Factors	Scores (Circle One)	Description/ Comments
1. Client and Extra-Contextual Factors		
1a.Client Factors		
1a1.Assessing client's **strengths**	0. Not at all 1. Minimal use 2. Effective use	
1a2. Assessing client's personal **motivation** to address psychosocial (mental health, physical, etc.) problem	0. Not at all 1. Minimal use 2. Effective use	
1a3. Assessing client's **current level of functioning** (e.g. domains of impairment vs. functioning such as family roles, work, school, community, etc.)	0. Not at all 1. Minimal use 2. Effective use	
1a4. Assessing client's **coping strategies**	0. Not at all 1. Minimal use 2. Effective use	
1a5. Assessing client's **recent life events**	0. Not at all 1. Minimal use 2. Effective use	

B

A. MICRO-SKILLS & THERAPIST[1] QUALITIES 0 - "Inappropriate" use refers to non-verbal and verbal communication that suggests hostility, negative judgment, blame, dismissive, ignoring the client, not paying attention, minimizing client's problems, etc. Examples may include no eye contact or too much eye contact, harsh tone of voice, only using yes/no questions, etc. 1 – "Socially acceptable" use refers to use of non-verbal and verbal communication in a way that is appropriate and normal for the cultural context when discussing psychosocial problems. 2 – "Therapeutically effective" use refers to use of techniques emphasized in psychosocial training and recognizable for a good therapist. Therapeutically effective approaches are comforting and show warmth, respect, attention, caring, and friendliness. Examples include a 'L'-shaped seat positioning, matching gestures to the client's level, a warm tone of voice, and interpreting client's feelings	Inappropriate [0]	Socially acceptable [1]	Therapeutically effective [2]	Not applicable [77]
Non-verbal communication [2] *(check the appropriate box)*				
A1. Eye contact	□	□	□	□
A2. Facial Expression	□	□	□	□
A3. Posture and positioning	□	□	□	□
A4. Gestures	□	□	□	□
A5. Use of Silence	□	□	□	□

C

ITEM 1. NON-VERBAL COMMUNICATION & ACTIVE LISTENING: EYE CONTACT, FACIAL EXPRESSION, BODY LANGUAGE & GESTURES

○ 1 NEEDS IMPROVEMENT = does not make any eye contact or stares at patient; shows anger; laughs at patient; mocks patient; turns away from patient; repeatedly interrupts patient; ignores patient; answers mobile phone without permission

○ 2 DONE PARTIALLY = does not consistently use body language to express interest; rarely makes eye contact; shows limited emotion; appears artificial

○ 3 DONE WELL = makes appropriate eye contact throughout interaction; smiles when appropriate; sits at appropriate angle from patient and leans in to show interest; use of 'uh-huh', 'hmm' or other culturally appropriate non-lexical utterances to signal interest

ITEM 2. VERBAL COMMUNICATION SKILLS: OPEN-ENDED QUESTIONS, SUMMARIZING & CLARIFYING STATEMENTS

○ 1 NEEDS IMPROVEMENT = uses mostly 'yes/no' questions, e.g., "Will you? Can you?"

○ 2 DONE PARTIALLY = uses open-ended questions but does not explore topics further or offer summaries for patient reflection

○ 3 DONE WELL = uses open-ended questions, summarizing and clarifies statements, e.g., "What happened? Tell me more."

ITEM 3. RAPPORT BUILDING & SELF-DISCLOSURE

○ 1 NEEDS IMPROVEMENT = clinician does not introduce him/herself or attempt to make the patient feel comfortable or clinician dominates the session talking about his/her own experiences

○ 2 DONE PARTIALLY = clinician introduces him/herself but does not help the patient feel comfortable through small talk and informal conversation or clinician disclosure but it is not related to patient experience or needs

○ 3 DONE WELL = clinician introduces him/herself, tries to make the patient feel comfortable and disclosure focuses on patient needs

Fig. 2. First (A), second (B), and final (C) iterations of scoring system for ENACT.

associated with an increase in response options (Miller, 1956; Anastasi, 1976; Krosnick, 1991). We viewed these concerns as particularly salient among non-specialist raters. We also chose to use qualifiers of skill performance (minimal use, effective use) over Likert ratings of agreement based on prior cross-cultural research on the limitations of the latter (Heine *et al.* 2002). Our prior work adapting psychiatric symptom scales in Nepal also highlighted problems with the use of Likert scales due to a cultural preference for agreement and acquiescence over disagreement ratings (Kohrt *et al.* 2011).

Pilot testing and scoring system revision

Government primary health workers who received mental health trainings participated in observed structured clinical evaluations (OSCEs), with standardized role-plays before and after the PRIME training. All role-plays in the study ranged from 15–20 min and covered a range of common clinical presentations including depression and self-harm. Role-play vignettes were developed by the study team (two Nepali therapists, Nepali study coordinator, Nepali training supervisor, and expatriate researchers—a psychiatrist and a psychologist—with greater than 10 years of experience working in Nepal). The vignettes were based on Nepali therapists' experiences with actual client interactions, for further information on vignette development, see (Kohrt *et al.* 2015). Vignettes were designed so that all domains on the ENACT scale would be applicable. The vignettes were then used to train Nepali psychosocial therapists to play the part of a standardized client.

OSCE role-plays were audio taped for transcription, translation, and qualitative data analysis (described below). After each rating, focus group discussions (FGDs) were conducted to qualitatively explore the feasibility, utility, and reliability of the scoring guidelines. FGDs included trainers, research coordinators, research assistants involved in trainings, the principal investigator, and a consultant. Process notes were recorded for analysis and to guide adaptation of the tool. The result of this process was a revision of the initial scoring system. Over multiple iterations, the tool was reduced from 49 to 18 items and the scoring system was revised.

Development of scoring criteria and an ENACT codebook

During the piloting phase, scoring criteria were modified and a preliminary codebook was developed. After finalizing the 18-item scale, an accompanying codebook was created. Data sources included observation of 30 primary care trainees enrolled in trainings, 26 standardized patient role-plays with non-specialists, and four expert sessions with actual clients (see Table 2). Of the standardized role-plays, 19 were post-training and seven were pre-training. All of the actual client sessions were conducted with an expert therapist who had more than 5 years of clinical experience. The audio recordings were transcribed in Nepali and translated into English. We employed a content analysis approach for the coding. In our prior review of common factors in existing instruments combined with common factor domain generation with Nepali counselors, we identified 115 common factor domains (Kohrt *et al.* 2015). These domains were used for the content analysis of the transcripts. This represented a deductive coding process because the individual codes had been established previously from common factors literature.

The qualitative data analysis software NVivo was used for all coding (QSR International, 2012). English transcripts were triple coded in NVivo by three study authors including one American psychiatrist, one Nepali psychosocial researcher, and one American graduate student. After collective development and revision of the codebook, transcripts were rated independently. Raters were not blinded to pre- *v.* post-training status of the transcript. Interrater reliability greater than 0.8-0 was required for each domain among the three coders.

Coding results were used to develop scoring criteria for each of the 18 items on the ENACT scale. Because this research was conducted within a task-shifting framework, criteria were particularly detailed so that they were easily comprehensible by non-specialists in similar lower-and-middle-income settings. Following criteria development, we consolidated results into an ENACT codebook for use by future expert raters. Codebook contents included detailed scoring criteria, along with culturally grounded examples of each of the 18 domains that exemplified a specific scoring level. Rationales for each example were provided. Below, we present the challenges encountered during development of the scoring system as documented through the process notes and standardized client role-plays.

Results

Trainees' perception of tool objectives: training and supervision v. research

One of the first challenges in implementing ENACT was the motivation of non-specialist providers to take part in standardized role-plays and conduct peer ratings. Without motivation to participate in role-plays, and complete the full version of the tool, the instrument would not contribute to quality improvement among peers. During the first application of the tool, one trainee, a health officer in charge of a primary care setting, initially refused to use the tool. This led us to revise how we introduced and explained ENACT in future trainings. We emphasized the tool's role in quality improvement for training and supervision to counter perceptions that it was only an

https://doi.org/10.1017/gmh.2015.21 Published online by Cambridge University Press

Table 2. *Non-specialist trainees completing pre- and post-training observed structured clinical evaluations with standardized clients*

Non-Specialist (NS)#	Age	Gender	Level of health training
NS #01	40	Male	Community medical auxiliary
NS #02	39	Male	Health assistant
NS #03	40	Male	Community medical auxiliary
NS #04	51	Male	Health assistant
NS #05	44	Male	Health assistant
NS #06	26	Male	Health assistant
NS #07	51	Male	Community medical auxiliary
NS #08	40	Male	Community medical auxiliary
NS #09	38	Male	Community medical auxiliary
NS #10	38	Male	Community medical auxiliary
NS #11	40	Male	Community medical auxiliary
NS #12	38	Female	Auxiliary health worker
NS #13	35	Female	Community medical auxiliary
NS #14	40	Male	Community medical auxiliary
NS #15	40	Male	Community medical auxiliary
NS #16	43	Male	Community medical auxiliary
NS #17	53	Male	Community medical auxiliary
NS #18	50	Male	Community medical auxiliary
NS #19	25	Female	Community medical auxiliary
NS #20	41	Male	Community medical auxiliary
NS #21	50	Male	Community medical auxiliary
NS #22	29	Female	Community medical auxiliary
NS #23	45	Male	Community medical auxiliary
NS #24	30	Female	Community medical auxiliary
NS #25	34	Male	Community medical auxiliary
NS #26	38	Male	Community medical auxiliary
NS #27	21	Male	Community medical auxiliary
NS #28	39	Male	Community medical auxiliary
NS #29	38	Male	Community medical auxiliary
NS #30	51	Male	Community medical auxiliary

instrument intended for research. For example, we changed our description of the tool by including relevant slides and examples of how it directly interfaced with common factors and psychosocial skills taught during the training. Another major change was to redesign the tool's response options, which were limited to the vague statements 'not at all', 'minimal use' and 'effective use'. After the third iteration, there were full descriptions of the three tiers of performance for each item (see Figure 2). Through this approach, the tool appeared more as a teaching tool that reinforced the ideal level of skill required to achieve mastery in each domain.

An example of using the tool to reinforce competency goals can be seen through ENACT item #9 'assessing recent life events and impact on psychosocial wellbeing'. Tier 1 was characterized by sessions in which the provider 'does not ask about triggering life events.' For a score of 2, the provider 'asks about life events but does not connect with current mental health needs'. For the highest score the provider 'asks about life events and discusses connection with current mental health needs'. By including a finer grain level of detail, peer raters could immediately see specific requirements for 'done well'. In prior scoring descriptions such as 'inappropriate' *v.* 'therapeutic', peers were not able to recall what the ideal skill application should be. The transcript below illustrates a provider who helps the client identify potential causes for insomnia and its relation to a buffalo loss, thus earning a score of 3 (of note, although scoring was assigned for each domain upon reviewing an entire session, the result below and subsequent quotes provide only a snapshot of how particular aspects of a domain were categorized into three competency tiers).

> ***Client:*** *I have been having problems sleeping for a long time. I don't enjoy working and have lost my appetite. It has been around eighteen days.*

https://doi.org/10.1017/gmh.2015.21 Published online by Cambridge University Press

> ***Provider:*** *It has already been eighteen days. What could be the reason for this?*
> ***Client:*** *I don't know.*
> ***Provider:*** *Why do you think it is hard to sleep?*
> ***Client:*** *Hmm, what are the reasons for that?*
> ***Provider:*** *Do you have any problems at home?*
> ***Client:*** *Yes, I do.*
> ***Provider:*** *And what are they?*
> ***Client:*** *I had a buffalo, which gave me lot of milk. It fell down and died.*
> ***Provider:*** *How long has it been?*
> ***Client:*** *It has been around twenty-five to thirty days. It died. I had invested around forty thousand (US$400) on it. Because of this, my life is sorrow and hardship.*
> [NS #03, 40-year-old male community medical auxiliary, post-training role-play]

Tool length

Trainers reported that completing the initial 49-item ENACT version was time-consuming for both trainers and trainees due to the high number of items. Further, they argued that it was difficult to keep all of the items in mind when observing structured clinical evaluations with standardized clients. Subsequently, one approach to item reduction was to employ the qualitative coding to group items together by clustering skills. The qualitative coders identified such clusters. For example, the item 'collaborative goal setting' and 'addressing client expectations', when done well were typically done together.

Therefore, for ENACT item #12, we categorized sessions in which the clinician dictated treatment goals and plans without client consultation as 1. For the second-tier, the provider asks about client goals, but does not discuss their feasibility or implementation. For the highest score, the provider and client jointly identify goals, discuss feasibility, and make action plans. Here, the provider works with the client to develop treatment goals and then to select a strategy for implementation. The provider highlights the client's agency in modifying goals and strategies (e.g. by increasing frequency of sessions), thus earning a score of 3.

> ***Provider:*** *So right now you are having a misunderstanding with your wife and have started drinking.*
> ***Client:*** *Yes. I don't want to go back home early as I do not want to argue with my wife and so I always end up getting drunk at my friend's place. I have started drinking a lot.*
> ***Provider:*** *So which issue do you think is important to address first?*
> ***Client:*** *I think if I have a good relationship with my wife, I can cut down my drinking.*
> ***Provider:*** *Okay, so then we will work on improving your marriage. We will meet every week, develop our plan of action and then review it and develop another plan in the next session. What do you think?*
> ***Client:*** *I think that is good. But, I think it would be better if we could meet sooner: maybe twice a week?*
> ***Provider:*** *Sure, we will meet every four days then.*
> [NS #05, 44-year-old male health assistant, post-training role-play]

Hierarchy of skills

One of the modifications to the tool that incorporated both the need to collapse items and the need to have the tool serve an educational function was to create a hierarchy of clinical skills within the scoring system. In the tool's initial version, there were 49 individual items with the associated scoring responses 'not at all', 'minimal use', and 'effective use'. Through qualitative coding, we found that numerous skills occurred together in a progression, where demonstration of higher order skill or a sequential skill routinely occurred after a lower order skill was displayed. For instance, there were initially separate items referring to exploration, interpretation, and normalization of feelings. However, interpretation and normalization occurred only if feelings were explored. Therefore, for ENACT item #4, we grouped these items together into one item within a three-tiered hierarchy, and we were able to capture emotionally judgmental and critical provider behaviors as well.

For ENACT item #4, a score of 1 refers to not asking about feelings or making judgmental, critical, or dismissive statements about emotions. A score of 2 categorizes statements where emotions were elicited but not normalized. A score of 3 refers to appropriate eliciting, exploration, and normalization of the client's emotional experience. In the example below, the provider helps the client explore and interpret her feelings and their associated contexts. The provider normalizes the happiness the client feels in her 'heart-mind' while present with her child compared with other activities. Heart-mind refers to the Nepali center of emotion, memory, and individuality (Kohrt & Harper, 2008). This combination illustrated below would be consistent with a score of 3.

> ***Provider:*** *May I know about your daily routine?*
> ***Client:*** *Yesterday morning, I woke up at around six. Then I washed my face. I didn't even eat anything.*
> ***Provider:*** *When you woke up, did you feel that your head was heavy? Were you sad or happy then?*

https://doi.org/10.1017/gmh.2015.21 Published online by Cambridge University Press

> ***Client:*** *My head was very heavy.*
> ***Provider:*** *After you washed your face, did you feel happy and fresh?*
> ***Client:*** *I didn't notice that.*
> ***Provider:*** *So continue, what did you do during the daytime?*
> ***Client:*** *I washed my face and drank tea. I have a child. My child woke up and I prepared to feed him milk. My heart-mind was happy then.*
> ***Provider:*** *So your heart-mind was happy when you were taking care of your child? What did you do after that?*
> ***Client:*** *After that, I came to the office at eight in the morning. I am at the office from eight to five. I forget these things while I am working.*
> ***Provider:*** *While you were busy in your office, did you feel happy or sad?*
> ***Client:*** *I don't remember, as I was too busy in office work. When I returned back, I enjoyed being with my child. After that I tried to sleep but couldn't at all.*
> ***Provider:*** *Okay, you say this is your problem. You did not get sleep. You feel happy when you are playing with your child, right?*
> ***Client:*** *A lot.*
> ***Provider:*** *Being mothers, we feel happy playing with our children. You became happy while you were playing with your child. Can you tell me anything besides that when you became happy yesterday?*
> [NS #12, 38-year-old female auxiliary health worker, post-training role-play]

Similarly, we were able to group four items on collaborative problem identification and solving into one item (ENACT item #15) with the three tiers reflecting sequential completion of steps required for effective problem solving.

Subjective v. objective ratings

Because of concerns regarding rating subjectivity, the first incarnation of the ENACT scale did not include Likert agreement responses. In prior work in Nepal, we found that the concept of agreement (e.g. totally agree *v.* somewhat agree) was seen as a personal judgment rather than an objective observation and therefore would interfere with making and sharing peer ratings. In addition, during piloting we found that both expert trainers and peer trainees had questions regarding objectively rating concepts such as 'promoting realistic hope for change' (ENACT item 13). Therefore, detailed scoring descriptions were required. The lowest tier included interactions where the provider does nothing to establish hope for change or, in the Nepali context, establishes unrealistic goals for improvement (e.g. telling a client with chronic psychosis that he/she will undoubtedly be cured in 3 months and can then stop medication forever). The second-tier score refers to referencing vague, positive outcomes, and promoting unrealistic hope. The highest tier is the development of hope that is reasonable, exemplified by discussing specific, context-appropriate treatment goals.

The account below includes an assessment of whether or not the client feels she can get better. However, the provider then makes a generalizing comment that fosters unrealistic expectations for change. Therefore, this account would be 2.

> ***Provider:*** *Do you think this is a situation that can get better, or do you believe it will always be this way?*
> ***Client:*** *I think this is just how life is and will be. I've been sad for a long time. I don't believe things can change.*
> ***Provider:*** *Well, you showed up for therapy today. Don't you think that means you might have some hope? If we are doing the right thing then what is there to be afraid of? I don't think your friends won't understand. Your friends are literate, have social status, and you are also in that group. If you don't want to stay away from your friends, then you can bring your friends too and we will discuss. You can tell them that if they too stay away from drinking then things will be better. It will be easier. Now we have to first make a plan. You have come here with a deep desire of quitting drinking. This is a plus point. Let us now make a plan and accordingly you can work on decreasing drinking. Then we can be successful. When can we meet again? Let's meet again and discuss so that I can help in making your effort successful.*
> ***Client:*** *I will come after a week. You talked about my friends. No matter what, they are still friends. I will try to talk with them once. Let's see what they say?*
> ***Provider:*** *So you will come after a week. Try to decrease your drinking level too. Try to control your heart-mind. Okay?*
> ***Client:*** *I will try.*
> ***Provider:*** *If you try, you will succeed. There is nothing that you cannot do.*
> [NS #20, 41-year-old male community medical auxiliary, post-training role-play]

Behavior counts

In attempt to promote objectivity, we attempted to use behavior counts for the different scoring levels. If a behavior was done 'X' number of times, the score was 2 'done partially', *v.* 'Y' number of times resulting in a score of 3 'done well'. For example, for ENACT

https://doi.org/10.1017/gmh.2015.21 Published online by Cambridge University Press

item 2 'verbal communication skills', the qualitative coders characterized Tier 1 ratings as those including fewer than two open-ended questions between provider and client, no summarizing statements, or no attempt at clarification with the client. Meeting any of these criteria led to a score of 1. Tier 2, was characterized by use of two or more open-ended questions, at least one summarizing statement, and at least one clarification statement. Tier 3 included meeting Tier 2 criteria plus greater than two summarizing statements. We chose to include these behavior counts because a similar approach was used with Uganda paraprofessional counselor evaluation (Kabura *et al.* 2005).

However, we found that both expert raters and participants had difficulty keeping exact behavioral counts in mind. Therefore, the response options were changed to reflect increasingly sophisticated communication skills moving from closed-ended questions (Tier 1) to open-ended questions (Tier 2) to both open-ended questions and summarizing statements (Tier 3), but without reference to the exact number of such occurrences.

The selection below is from a health worker trainee in a standardized role-play. The selection is taken from a role-play that scored 3 based on repeated use of summarizing, open-ended questions, and clarifying statements.

> ***Provider:*** *I will go back in time. How long has it been since that you got married?*
> ***Client:*** *It has been around 15–16 years.*
> ***Provider:*** *Oh, it has already been 15–16 years. That means you have two children, a husband, and have been married for 16 years. Can you tell me how your relationship is with your husband? Your relationship with your husband is now nearly 2 decades long. But still, if you can say something about your relationship with your husband.*
> ***Client:*** *It is not that good. It is just okay.*
> ***Provider:*** *You said it is not that good. Can you explain how it is?*
> ***Client:*** *He does not listen to me nor take my advice. He dominates me a lot.*
> ***Provider:*** *That means even though you have been together for 16 years after marriage, you don't share a lot with your husband. Because of that, you are feeling that your husband is not helping and loving you. How is your husband's relationship with people outside your home? Do you have any doubts?*
> [NS #24, 30-year-old female community medical auxiliary, post-training role-play]

Use of 'not applicable' rating option

In the initial scoring system design, 'not applicable' was included as a response option. However, we found that this response was often selected in error. For example, when trainees failed to rate item #18 'safety and suicidality', peers would rate the item as not applicable rather than scoring as 'needs improvement', the lowest tier. Moreover, a significant amount of time in trainings and scoring was also lost due to debates about what was considered 'not applicable'. Therefore, we created role-plays in which all 18 items of the ENACT were applicable. Trainees were then instructed to place a score of 'needs improvement' if the action was not done.

Social desirability and response set options

From the outset, three response options were used because of our intention to include the fewest number of response options required to capture adequate variance and change over time. In the first iteration of the tool, the three response options were 0 'not at all', 1 'minimal use', 2 'effective use'. This structure became difficult to employ because it only was appropriate for items where we were focusing on the presence of a behavior. It, however, was not meaningful when one intended to capture behaviors that should be avoided. Therefore, in the second incarnation, the scoring was 0 'inappropriate', 1 'socially acceptable', 2 'therapeutically effective', and 77 'not applicable'. However, participants reported that this approach was too subjective. Based on caste/ethnicity, age, gender, occupation, and other demographic markers, what is considered 'socially acceptable' and 'inappropriate' may vary. In the third format, we employed 0 needs improvement, 1 done partially, and 2 done well. Although the wording was considered acceptable, we initially found very few 0 responses. When non-specialists were asked about this, they reported that it was unacceptable to rate their peers as 0. We then maintained the text descriptions but changed the numbers to 1–2–3 instead of 0-1-2. By changing 0–1 and shifting the scale upwards, we found an increased use of the lowest score option 1.

Discussion

The qualitative process evaluation during development of the scoring system for ENACT revealed challenges in the creation of peer rating tools for non-specialist providers. The challenges can be grouped into five domains: (1) balance of training and supervision objectives with research objectives; (2) burden for peer raters due to number of scale items, number of response options, and use of behavioral counts; (3) incorporating a hierarchy of clinical skills; (4) objective *v.* subjective aspects of rating; and, (5) social desirability when rating peers.

https://doi.org/10.1017/gmh.2015.21 Published online by Cambridge University Press

Five challenges for non-specialist peer rating tools

(1) *To facilitate adoption of tools for quality improve purposes, the value of tools for training and supervision needs to be clarified beyond use as research instruments*. ENACT needed to represent a teaching aid. This led to changes regarding details for each item specifying what was meant by 'done well' *v.* 'needs improvement'. The CTS-R is a one example of this in the current literature, as it provides information and key features of each skill and descriptions for each scoring level (Blackburn *et al.* 2001). Tools such as Counselor Rating Form-Shortened version, CRF-S (Corrigan & Schmidt, 1983) that use general responses such as 'very' to 'not very' may have less appeal for peer non-specialist ratings because the direct relevance to clinical improvement and the educational value of the tool is more opaque.

(2) *Regarding scalability, tools need to be user-friendly, which includes limiting the number of items and the number of response set options, as well as avoiding taxing concentration with a range of behavioral counts*. Tools with nine Likert response options or 100 items are less amenable to non-specialist peers who are burdened with a range of other healthcare or social work duties. In Figure 3, we compare ENACT with other therapist rating tools based on the total number of scoring decisions (total number of items X number of response options per item) as approximation of the burden on cognition and time. At the extreme upper end is the HIM-G version with over 400 response decisions, exemplifying a tool that gathers a tremendous amount of data but is limited to use by experts. The Patient Session Questionnaire and Working Alliance Inventory-Revised have over 200 response decisions, thus also burdensome for peer non-specialists. A number of existing tools have approximately 90 response decisions (CRF-S version, Helping Alliance Questionnaire, CTS-R). ENACT was designed to have a minimal burden on peers with 54 response decisions (18 items with three response options per item). Despite successful use of behavior counts in a Ugandan study (Kabura *et al.* 2005), participants in Nepal reported difficulty keeping track of counts when making live ratings. We also dropped 'not applicable' because the time and confusion related to when it should be invoked.

(3) *One component of optimizing training and supervision utility as well as minimizing the number of items is to use a scoring system that captures skills hierarchies*. The CTS-R is one of the few tools that incorporates this. Other tools using Likert agreement endorsements per items do not distinguish between degree to which a simple skill was done well and the level of complexity with which that skill was demonstrated.

(4) *When rating peers and working with non-specialists, it was important to minimize perception that scoring represents subjective personal appraisals or opinions*. Likert agreement scales were not selected and detailed descriptions were needed so that peers could point toward specific behaviors that justified a score. Moreover, by anchoring each score with a description, it alleviates anxiety and time burden related to distinguish between scores such as a '5' and '6' on a nine-item Likert scale that has a limited

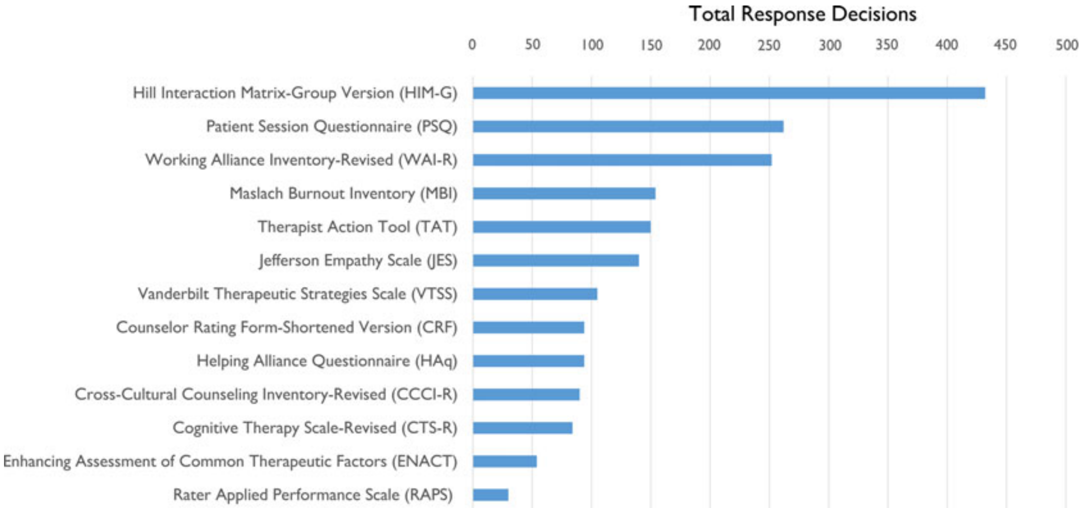


Fig. 3. Total response decisions for therapist rating tools. Total response decisions refer to number of items multiplied by the number of response options per item.

number of anchors. This may be especially important in cultural context where group consensus is emphasized (Heine *et al.* 2002).

(5) *When asking peers to rate one another, social desirability biases need to be considered.* Through small changes such as using 1 as the lowest score with a description of 'needs improvement', we were able to increase use of this response option compared with when the lowest score was 0 described as 'not at all' or 'inappropriate'.

Taken together, these issues engage with global mental health debates regarding the tension between 'excellence' and 'relevance', with researchers focusing on detailed, elaborate information, and practitioners interested in issues of application and 'good enough' services (Tol *et al.* 2012). ENACT was developed to provide information to contribute to excellence in research, but with a scope and burden that would enable it to be relevant for non-specialists to put into practice, thus increasing the likelihood of scalability.

Application of ENACT scoring system

ENACT's scoring system lends itself to a number of applications. The scoring system can be used to guide decision-making. Tier 1 reflected the need for significant improvement and remediation through training and supervision. For situations where minimum competence is required, e.g. participating in a research trial, then persons achieving only or mostly Tier 1 competency (post training) across many domains may not be selected. Tier 2 demonstrates partial competence of a common factor. Program designers could anticipate Tier 2 competencies to be achieved in most domains after training has been completed but before a protracted period of supervision begins. Following supervision, Tier 3 would be a reasonable expectation. Tier 3 is the optimal level for basic competence in a common factor. After training and supervision for a pre-determined period (e.g. 3–6 months), non-specialists would be expected to achieve tier 3 on most domains.

In addition, ENACT can be used to identify specific competencies which could be considered absolute minimal requirements for involvement in task-sharing care delivery. For example, Murray *et al.* (2014*b*) identified safety planning for risk of harm (self-harm, intimate partner violence, and child maltreatment) as key domains in task-sharing which are often neglected. Therefore, third-tier competency on ENACT item #18 could be a minimum requirement for practice. Similarly, certain task-sharing programs may emphasize specific common factors more than others do based on their theory of change models and implementation. For example, the 'Friendship Bench' program in Zimbabwe relies strongly upon problem solving (Chibanda *et al.* 2011), and thus, achievement of third-tier competency would be a minimum requirement on ENACT item #15.

From an implementation science and program evaluation perspective, the percentage of non-specialists achieving competency on each item can be used to highlight areas for change. For example, if there are certain items where most non-specialists achieve only Tier 1 following training, training content would need to be improved in that domain. If, after 6 months of supervision most non-specialists were still at Tier 2, then supervision intensity and strategy should be adjusted. Tools such as ENACT can be used to enhance apprentice-based supervision strategies that focus on individual provider needs and skill development (Murray *et al.* 2011); for example, ENACT can be used to identify which common factors are at the lower competency tiers and select them for structured remediation.

The applications of ENACT are not limited to LMICs. ENACT and similar endeavors are important for high-income countries (HIC), such as countries in North America and Europe, where there is also significant burden of unaddressed mental health needs. Gaps in services are especially pronounced among racial and ethnic minority groups (Alegría *et al.* 2008). Tools built for cross-cultural use could be ideal for non-specialists from these communities delivering mental health services. For many populations in HIC especially minority ethnic groups, primary care and community-based providers are more likely to be delivering mental health services (Wang *et al.* 2006). These providers may be preferred because of financial barriers, lack of transportation, stigma associated with mental health professionals, and other cultural issues in explanatory models and help-seeking behavior that limit use of specialty mental health services (Scheppers *et al.* 2006).

For non-mental health specialists in HIC to adopt quality improvement tools for mental health, tools need to address the five challenges identified in our results presented here. Based on our review of existing 56 common factors tools developed in high-income settings (Kohrt *et al.* 2015), the majority of tools do not adequately counter these potential barriers. Therefore, ENACT and similarly developed tools should be explored for use with non-specialist providers in high-income settings. Tools such as ENACT can be useful to evaluate competence among primary care workers, community health workers, peer helpers, cultural brokers, refugee resettlement workers, teachers, and paraprofessionals. This raises an important distinction between tools to assess cultural competence of expert therapist (of which there are many examples) and tools that are cross-culturally appropriate for rating

https://doi.org/10.1017/gmh.2015.21 Published online by Cambridge University Press

common factors among diverse care providers with limited mental health expertise. ENACT fills a gap in this latter arena.

Moreover, because of the debates regarding use of non-specialists in HIC focuses on concerns about quality of care (Robiner, 2006; Unützer *et al.* 2006; Montgomery *et al.* 2010; Fuller *et al.* 2011), ENACT-type tools can be used to objectively evaluate performance for certification or employment. Objective demonstration of achieving and maintaining competency may be helpful to assure that these cadres of workers can be appropriately compensated for providing services.

Limitations

The current results are limited to formative work done in Nepal. Additional sites and types of programs will need to adapt competency tiers based on their setting and implementation context. Although a three-tiered response scale appeared to have feasibility and utility from a qualitative perspective in Nepal, quantitative psychometric studies are needed to evaluate distribution of scores and pre-/post-training changes. Of note, initial qualitative process evaluation of the ENACT tool in Liberia suggested that a four-tiered response scale may be needed to prevent ceiling effects and enable non-specialists to track improvement over longer periods of time while providing mental health services (that is, to have a higher level to strive for as opposed to communicating a message that a third-tier score is terminally sufficient). Another component to incentive further learning through peer non-specialist ratings is the use of treatment-specific scales appropriate for non-specialists that address the five challenges raised here.

Important next steps will include evaluating the ability to differentiate among the tiers when rating the transcripts in Nepali rather than English, and when rating direct observation or video files. Lack of blinding to pre- and post-training status is a major limitation for the quantitative analysis of the scores. The lack of blinding likely biases raters to inflate scores after training, and potentially depress scores prior to training because of unconscious biases regarding what participants should be able to do at different stages. Similarly, there may be social desirability bias for inflated post-training scores because raters want scores to reflect positively on the trainers. Therefore, blinding is currently being used for coding comparisons between pre- *v.* post-training competence. Similarly, for other training and supervision studies, objective evaluations of competence should include blinded ENACT ratings in addition to those completed during trainings. Ultimately, the utility of the scale and the proposed three-tier rating categories needs to be assessed against client outcomes. This work is currently underway through PRIME in Nepal. The process documented here represents the crucial preliminary steps toward analyzing the pathway from training to client well-being in task-sharing initiatives.

Conclusion

With increasing attention to task-sharing in global mental health, there is a need to assure that non-specialists achieve and maintain a minimum standard of competence and quality in order to adhere to evidence-based practices, strive for positive client outcomes, and reduce risk of harm. Lack of human resources in the form of mental health experts is a challenge not only for service delivery but also for training and supervision. Task-sharing aspect of training and supervision so that non-specialists can improve quality among their peers is one strategy to improve implementation and dissemination. Development of tools for non-specialists to use with peers can facilitate this process. We identified five challenges for development and use of non-specialist peer rating tools and proposed approaches to address these challenges as demonstrated through development of the ENACT scale. Ultimately, these challenges and the need for tools such as ENACT are not limited to low-resource setting. These are needs in high-resource settings where task-sharing is occurring as well. Future work is required for development and implementation of user-friendly, value added peer-rating system across settings varied by culture, income, and mental health needs.

Supplementary material

For supplementary material accompanying this paper visit http://dx.doi.org/10.1017/gmh.2015.21.

Acknowledgements

The authors thank Matthew Burkey, Michael Compton, James Griffith, Revathi Krishna, Nagendra Luitel, Crick Lund, Vikram Patel, Pragya Shrestha, Daisy Singla, and Lawrence Wissow for their valuable discussions during the tool development process.

Financial Support

The research was conducted through the South Asia Hub for Advocacy, Research, and Education for Mental Health (SHARE) funded by NIMH (U19MH 095687-01S1), Reducing Barriers to Mental Health

Task Sharing funded by NIMH (K01MH104310-01), and the Programme to Improve Mental Health Care (PRIME) funded by the Department for International Development (DFID). This document is an output from the PRIME Research Programme Consortium, funded by UK aid from the UK government. However, the views expressed do not necessarily reflect the UK government's official policies. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. The authors had full control of all primary data.

Declaration of Interest

None.

Ethical Standards

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation, and with the Helsinki Declaration of 1975, as revised in 2008. The study protocol was approved by the Nepal Health Research Council, Kathmandu, Nepal.

References

Abas M, Baingana F, Broadhead J, Iacoponi E, Vanderpyl J (2003). Common mental disorders and primary health care: current practice in low-income countries. *Harvard Review of Psychiatry* **11**, 166–173.

Alegría M, Chatterji P, Wells K, Cao Z, Chen C-n, Takeuchi D, Meng X-L (2008). Disparity in depression treatment among racial and ethnic minority populations in the United States. *Psychiatric Services* **59**, 1264–1272.

Anastasi A (1976). *Psychological Testing*. 4th edn. Macmillan: New York.

Biernat M, Manis M, Nelson TE (1991). Stereotypes and standards of judgment. *Journal of Personality and Social Psychology* **60**, 485.

Blackburn IM, James IA, Milne DL, Baker C, Standart S, Garland A, Reichelt FK (2001). The revised cognitive therapy scale (CTS-R): psychometric properties. *Behavioural and Cognitive Psychotherapy* **29**, 431–446.

Cahill J, Barkham M, Hardy G, Gilbody S, Richards D, Bower P, Connell J (2008). A review and critical appraisal of measures of therapist-patient interactions in mental health settings. *Health Technology Assessment* **12**, iii, ix–47.

Chibanda D, Mesu P, Kajawu L, Cowan F, Araya R, Abas M (2011). Problem-solving therapy for depression and common mental disorders in Zimbabwe: piloting a task-shifting primary mental health care intervention in a population with a high prevalence of people living with HIV. *BMC Public Health* **11**, 828.

Corrigan JD, Schmidt LD (1983). Development and validation of revisions in the counselor rating form. *Journal of Counseling Psychology* **30**, 64–75.

Fairburn CG, Cooper Z (2011). Therapist competence, therapy quality, and therapist training. *Behaviour Research and Therapy* **49**, 373–378.

Frank JD, Frank JB (1991). *Persuasion and Healing: A Comparative Study of Psychotherapy*. 3rd edn. Johns Hopkins University Press: Baltimore.

Fuller J, Perkins D, Parker S, Holdsworth L, Kelly B, Roberts R, Fragar L (2011). Building effective service linkages in primary mental health care: a narrative review part 2. *BMC Health Services Research* **11**, 66.

Garfield SL (1973). Basic ingredients or common factors in psychotherapy? *Journal of Consulting and Clinical Psychology* **41**, 9–12.

Greenberg RP (2004). Essential ingredients for successful psychotherapy: effect of common factors. In *The Art and Science of Brief Psychotherapies: A Practitioner's Guide* (Eds M. J. Dewan, B. N. Steenbarger and R. P. Greenberg), pp. 231–241. American Psychiatric Publishing, Inc.: Arlington, Virginia, USA.

Heine SJ, Lehman DR, Peng K, Greenholtz J (2002). What's wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect. *Journal of Personality and Social Psychology* **82**, 903–918.

Henry WP, Schacht TE, Strupp HH, Butler SF, Binder JL (1993). Effects of training in time-limited dynamic psychotherapy: mediators of therapists' responses to training. *Journal of Consulting & Clinical Psychology* **61**, 441–447.

Hill CE, Gormally J (1977). Effects of reflection, restatement, probe, and nonverbal behaviors on client affect. *Journal of Counseling Psychology* **24**, 92–97.

Hoyt MF, Marmar CR, Horowitz MJ, Alvarez WF (1981). The therapist action scale and the patient action scale – instruments for the assessment of activities during dynamic psychotherapy. *Psychotherapy-Theory Research and Practice* **18**, 109–116.

Ivey AE, Authier J (1978). *Microcounseling: Innovations in Interviewing, Counseling, Psychotherapy, and Psychoeducation*. Charles C. Thomas: Springfield, Illinois, USA.

Jenkins R, Othieno C, Okeyo S, Kaseje D, Aruwa J, Oyugi H, Kauye F (2013). Short structured general mental health in service training programme in Kenya improves patient health and social outcomes but not detection of mental health problems-a pragmatic cluster randomised controlled trial. *International Journal of Mental Health Systems* **7**, 25.

Jordans M, Luitel N, Tomlinson M, Komproe I (2013). Setting priorities for mental health care in Nepal: a formative study. *BMC Psychiatry* **13**, 332.

Jordans MJD, Luitel NP, Pokhrel P, Patel V (2015). Development and pilot testing of a mental healthcare plan in Nepal. *The British Journal of Psychiatry*. Supplement, S1–S8. doi: 10.1192/bjp.bp.114.153718.

Kabura P, Fleming LM, Tobin DJ (2005). Microcounseling skills training for informal helpers in Uganda. *International Journal of Social Psychiatry* **51**, 63–70.

Kakuma R, Minas H, van Ginneken N, Dal Poz MR, Desiraju K, Morris JE, Scheffler RM (2011). Human resources for mental health care: current situation and strategies for action. *The Lancet* **378**, 1654–1663.

Karson M, Fox J (2010). Common skills that underlie the common factors of successful psychotherapy. *American Journal of Psychotherapy* **64**, 269–281.

Kohrt BA, Harper I (2008). Navigating diagnoses: understanding mind-body relations, mental health, and stigma in Nepal. *Culture, Medicine & Psychiatry* **32**, 462–491.

Kohrt BA, Jordans MJ, Tol WA, Luitel NP, Maharjan SM, Upadhaya N (2011). Validation of cross-cultural child mental health and psychosocial research instruments: adapting the depression self-rating scale and child PTSD symptom scale in Nepal. *BMC Psychiatry* **11**, e127.

Kohrt BA, Jordans MJD, Rai S, Shrestha P, Luitel NP, Ramaiya M, 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**, 11–21.

Krosnick JA (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology* **5**, 213–236.

Lafromboise TD, Coleman HLK, Hernandez A (1991). Development and factor structure of the cross-cultural counseling inventory revised. *Professional Psychology-Research and Practice* **22**, 380–388.

Lipsitz J, Kobak K, Feiger A, Sikich D, Moroz G, Engelhard A (2004). The rater applied performance scale: development and reliability. *Psychiatry Research* **127**, 147–155.

Luborsky L, Barber JP, Siqueland L, Johnson S, Najavits LM, Frank A, Daley D (1996). The revised helping alliance questionnaire (HAq-II): psychometric properties. *The Journal of Psychotherapy Practice and Research* **5**, 260–271.

Lund C, Tomlinson M, De Silva M, Fekadu A, Shidhaye R, Jordans M, Patel V (2012). PRIME: a programme to reduce the treatment gap for mental disorders in five low- and middle-income countries. *PLoS Medicine/Public Library of Science* **9**, e1001359.

Makanjuola V, Doku V, Jenkins R, Gureje O (2012). Monitoring and evaluation of the activities of trainees in the 'training of trainers' workshop at Ibadan, south-west Nigeria. *Mental Health in Family Medicine* **9**, 25–32.

Margison FR, McGrath G, Barkham M, Clark JM, Audin K, Connell J, Evans C (2000). Measurement and psychotherapy: evidence-based practice and practice-based evidence. *The British Journal of Psychiatry* **177**, 123–130.

Maslach C, Jackson SE, Leiter MP (1996). *Maslach Burnout Inventory Manual*. 3rd edn. Consulting Psychologists Press, Inc.: Palo Alto, CA.

Miller GA (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review* **63**, 81–97.

Montgomery EC, Kunik ME, Wilson N, Stanley MA, Weiss B (2010). Can paraprofessionals deliver cognitive-behavioral therapy to treat anxiety and depressive symptoms? *Bulletin of the Menninger Clinic* **74**, 45–62.

Murray LK, Dorsey S, Bolton P, Jordans MJ, Rahman A, Bass J, Verdeli H (2011). Building capacity in mental health interventions in low resource countries: an apprenticeship model for training local providers. *International Journal of Mental Health Systems* **5**, 30.

Murray LK, Tol W, Jordans M, Sabir G, Amin AM, Bolton P, Thornicroft G (2014*a*). Dissemination and implementation of evidence based, mental health interventions in post conflict, low resource settings. *Intervention* **12**, 94–112.

Murray LKP, Skavenski SMSWMPH, Bass JP, Wilcox HP, Bolton PMMPH, Imasiku MP, Mayeya JMPH (2014*b*). Implementing evidence-based mental health care in low-resource settings: a focus on safety planning procedures. *Journal of Cognitive Psychotherapy* **28**, 168–185.

O'Malley SS, Suh CS, Strupp HH (1983). The Vanderbilt psychotherapy process scale: a report on the scale development and a process-outcome study. *Journal of Consulting and Clinical Psychology* **51**, 581–586.

Patel V (2009). The future of psychiatry in low- and middle-income countries. *Psychological Medicine* **39**, 1759–1762.

Patel V (2012). Global mental health: from science to action. *Harvard Review of Psychiatry* **20**, 6–12.

Patel V, Weobong B, Nadkarni A, Weiss H, Anand A, Naik S, Kirkwood B (2014). The effectiveness and cost-effectiveness of lay counsellor-delivered psychological treatments for harmful and dependent drinking and moderate to severe depression in primary care in India: PREMIUM study protocol for randomized controlled trials. *Trials* **15**, 101.

QSR International (2012). NVivo qualitative data analysis software (Version 10). QSR International Pty Ltd.: Doncaster, Australia.

Robiner WN (2006). The mental health professions: workforce supply and demand, issues, and challenges. *Clinical Psychology Review* **26**, 600–625.

Sadik S, Abdulrahman S, Bradley M, Jenkins R (2011). Integrating mental health into primary health care in Iraq. *Mental Health in Family Medicine* **8**, 39–49.

Samstag LW, Batchelder ST, Muran JC, Safran JD, Winston A (1998). Early identification of treatment failures in short-term psychotherapy: an assessment of therapeutic alliance and interpersonal behavior. *Journal of Psychotherapy Practice & Research* **7**, 126–143.

Scheppers E, van Dongen E, Dekker J, Geertzen J, Dekker J (2006). Potential barriers to the use of health services among ethnic minorities: a review. *Family Practice* **23**, 325–348.

Schindler L, Hohenberger-Sieber E, Hahlweg K (1989). Observing client—therapist interaction in behaviour therapy: development and first application of an observational system. *British Journal of Clinical Psychology* **28**, 213–226.

Singla DR, Weobong B, Nadkarni A, Chowdhary N, Shinde S, Anand A, Weiss H (2014). Improving the scalability of psychological treatments in developing countries: an evaluation of lay therapist led therapy quality in Goa, India. *Behaviour Research and Therapy* **60**, 53–59.

Sparks JA, Duncan BL, Miller SD (2008). Common factors in psychotherapy. In *Twenty-First Century Psychotherapies: Contemporary Approaches to Theory and Practice* (ed. J. L. Lebow), pp. 453–497. John Wiley & Sons Inc: Hoboken, NJ, US.

Suh DH, Hong JS, Lee DH, Gonnella JS, Hojat M (2012). The Jefferson scale of physician empathy: a preliminary

https://doi.org/10.1017/gmh.2015.21 Published online by Cambridge University Press

psychometric study and group comparisons in Korean physicians. *Medical Teacher* **34**, E464–E468.

Tol WA, Patel V, Tomlinson M, Baingana F, Galappatti A, Silove D, Panter-Brick C (2012). Relevance or excellence? Setting research priorities for mental health and psychosocial support in humanitarian settings. *Harvard Review of Psychiatry* **20**, 25–36.

Unützer J, Schoenbaum M, Druss BG, Katon WJ (2006). Transforming mental health care at the interface with general medicine: report for the Presidents Commission. *Psychiatric Services* **57**, 37–47.

van Ginneken N, Tharyan P, Lewin S, Rao GN, Meera S, Pian J, Patel V (2013). Non-specialist health worker interventions for the care of mental, neurological and substance-abuse disorders in low-and middle-income countries. *The Cochrane Database of Systematic Reviews* **11**, CD009149.

Wampold BE (2011). The research evidence for common factors models: a historically situated perspective. In *The Heart and Soul of Change: Delivering What Works in Therapy*, 2nd edn (ed. B. L. Duncan, S. D. Miller, B. E. Wampold and M. A. Hubble), pp. 49–82. American Psychological Association: Washington, DC.

Wang PS, Demler O, Olfson M, Pincus HA, Wells KB, Kessler RC (2006). Changing profiles of service sectors used for mental health care in the United States. *American Journal of Psychiatry* **163**, 1187–1198.

WHO (2008). *Task Shifting: Rational Redistribution of Tasks Among Health Workforce Teams: Global Recommendations and Guidelines*. WHO Press: Geneva.

WHO (2010). *mhGAP Intervention Guide for Mental, Neurological and Substance-Use Disorders in Non-Specialized Health Settings: Mental Health Gap Action Programme (mhGAP)* (pp. 83). WHO Press: Geneva.

https://doi.org/10.1017/gmh.2015.21 Published online by Cambridge University Press