RUNNING HEAD: PEER SUPERVISION IN A PEER-DELIVERED INTERVENTION

Peer supervision for assuring the quality of non-specialist provider delivered psychological intervention: Lessons from a trial for perinatal depression in Goa, India

Daisy R. Singla[1-2], Clarissa Ratjen[1], Revathi N. Krishna[3-4], Daniela C. Fuhr[5], Vikram Patel[3, 7-8]

[1]Department of Psychiatry, Sinai Health System, Toronto, Canada

[2]Department of Psychiatry, University of Toronto, Toronto, Canada

[3]Sangath, Goa, India

[4]Monash University Accident Research Centre (MUARC), Monash University, Melbourne, Australia

[5]Department of Health Services Research and Policy, London School of Tropical Medicine and Hygiene, London, United Kingdom

[6]Department of Global Health and Social Medicine, Harvard Medical School, Cambridge, USA

[7]Department of Global Health and Population, Harvard TH Chan School of Public Health, Cambridge, USA

***Corresponding Author:** **Daisy R. Singla**
Department of Psychiatry
University of Toronto, Sinai Health System
600 University Ave, Rm 914A
Toronto, ON, Canada M5G 1X5
Email**:** daisy.singla@utoronto.ca

ABSTRACT

Background. The aims of the current study were three-fold: i) to estimate the reliability and predictive validity of a therapy quality measure for use by peers; ii) to assess the extent to which peer delivery agents could be trained to evaluate their peers' counsellors as reliably as experts; and iii) to identify barriers and facilitators of several implemented models of peer supervision.

Methods. 26 peers (called 'Sakhis' in the study context), with no previous experience or formal training in mental health care delivery, were trained by experts to deliver the Thinking Healthy Program Peer-delivered (THPP) and conduct peer-led supervision. Using the Therapy Quality Scale (TQS)—an 18 item Likert scale (0-2) measuring both general and treatment-specific skills—both peers and experts independently rated 167 individual sessions to estimate: a) the psychometric properties of TQS; and b) the mean difference between peer and expert TQS ratings; these data were analyzed using SAS 9.3. This was complemented with qualitative data (two rounds of in-depth interviews with four experts and focus group discussions with all Sakhis) which were analyzed using framework analysis.

Results. We observed good internal consistency on TQS ratings among expert (α=0.814) and Sakhis (α=0.843) and good to excellent scores of inter-rater reliability among experts (ICC=0.779) and Sakhis (ICC=0.714). TQS ratings were not significantly related to patient depressive symptoms at 6-months post-child birth but were significantly related to patient activation scores (r=0.375, p<0.01 for treatment-specific skills and 0.313, p<0.01 for general skills) at 3-months post-child birth, which in turn were significantly related to depressive symptom scores at 6-months post-child birth (r=-0.455, p<0.001), highlighting a potential temporal pathway between therapy quality, patient behaviours and patient outcomes. Following additional training and with growing expertise, Sakhis eventually evaluated their peers' counselling sessions as reliably as experts—demonstrating no significant mean differences on general (t=-0.42, p>0.05) or treatment-specific (t=-1.44, p>0.05) subscale scores. Qualitative findings were also consistent between experts and Sakhis: barriers included peers' initial difficulties in rating the TQS and leading supervision which declined over time. Most Sakhis and experts reported the benefits of using a structured scale to rate therapy quality which in turn facilitated consistent and relevant feedback and motivation to ultimately improve Sakhis' counselling skills. In addition, most Sakhis and experts found that peer supervision methods were acceptable and feasible, particularly when linked to financial incentives and expert supervisor.

Conclusion. With time, non-specialist or lay providers can be trained to implement peer supervision and assess therapy quality as reliably as experts using a psychometrically-sound measure. However, peer supervision with experts was more preferred than peer supervision without experts to facilitate structured, reliable feedback. Additional studies are required to address this challenge and test solutions to facilitate the dissemination of non-specialist delivered psychosocial interventions at a global level.

Keywords. therapy quality, depression, behavioral activation, non-specialist providers, peer supervision

INTRODUCTION

Worldwide, robust evidence exists to demonstrate the effectiveness of psychological treatments delivered by non-specialist providers (NSPs; Hoeft, Fortney, Patel, & Unützer, 2016; Singla et al., 2017). Despite their effectiveness, psychological treatments are poorly disseminated. This is partly due to the limited number of available mental health professionals who can train novice clinicians with high fidelity (Herschell, Kolko, Baumann, & Davis, 2010; Kazdin, 2017) and the way in which they are supervised (Fairburn & Patel, 2017). Supervision is considered a key pedagogical and quality assurance tool in treatment delivery, and known to influence the therapist in training (Saxon, Barkham, Foster, & Parry, 2016; Saxon, Firth, & Barkham, 2016), the therapeutic process (Ladany, Ellis, & Friedlander, 1999) and patient outcomes (Fairburn & Cooper, 2011; Watkins Jr, 2011). However, supervision typically involves practicing the treatment under the direction of someone experienced in delivering that treatment (Herschell et al., 2010). This is also the case in non-specialist delivered interventions which, while effective (Singla et al., 2017; Van Ginneken et al., 2013), rely on experts—typically a mental health professional—to train and supervise NSPs.

This bottleneck could be potentially addressed by the task-sharing of supervision to peers, while using standardized tools with robust psychometric properties. For example, a recent study from India found that lay counsellors were able to assess therapy quality as reliably as expert supervisors for psychological treatments for depression and harmful drinking (Singla et al., 2014). Furthermore, this peer supervision model was found to be both acceptable and feasible by lay counsellors. Peer supervision may also be advantageous because it encourages therapists in training to draw upon others' experiences (Golia & McGovern, 2013) and take active roles in assisting one another including the alleviation of stress, anxiety and feelings of anxiety (Yeh et al., 2008). In Uganda, structured peer supervision using checklists to monitor one's session delivery was perceived as advantageous by both trainee and supervisors because it offered concrete feedback to the delivering therapist (Singla & Kumbakumba, 2015). It remains unclear, however, whether these and other examples of measurement-based peer supervision within a task-sharing context can successfully continue without an expert present.

The current study examined the acceptability, feasibility and accuracy of measurement-based peer supervision in the SHARE—the **S**outh Asian **H**ub for **A**dvocacy, **R**esearch and **E**ducation on Mental Health—trial in India (Fuhr et al., 2019). The goal of the trial was to adapt the Thinking Healthy Program (THP; Rahman, Malik, Sikander, Roberts, & Creed, 2008), a psychological treatment for perinatal depression, for delivery by peers (called the Thinking Healthy Program, Peer-Delivered or THPP). The THP was originally developed and evaluated in Pakistan (Rahman et al., 2008) and is recommended by the World Health Organization for the treatment of perinatal depression in low-resource settings (http://www.who.int/mental_health/maternal-child/thinking healthy/en/). Peers—referred to "*Sakhis*" (meaning 'friend') were mothers living in the same or neighbouring community as mothers participating in the intervention. THPP in Goa, India (hereafter referred to as THPP-India; Sikander et al., 2015) trial found THPP to be more effective and cost-effective than enhanced usual care alone Peers were also found to be an acceptable and feasible delivery agent (Singla et al., 2014).

To our knowledge, few studies have systematically evaluated the process and reliability of peer-supervision, in particular in the context of task-sharing for mental health care. In addition, no study to our knowledge has examined the acceptability and feasibility of peer supervision with and without an expert present in a task-sharing context. We developed a tool for this purpose called the Therapy Quality Scale (TQS) and used this to rate the quality of therapy sessions. The aims of the study reported in this paper were:

1) To describe the psychometric properties of the Therapy Quality Scale, including its reliability (internal consistency, test re-test and inter-rater reliability) and validity (predictive validity in relation to patient activation at 3-months post-child birth and patient outcomes at 6-months post-childbirth).
2) To examine the accuracy of peer ratings as compared to expert supervisors, i.e., the extent to which Sakhis could be trained to evaluate their peers' counselling as reliably as experts
3) To identify the perceived barriers and facilitators among Sakhis and their supervisors related to the implementation of peer supervision, including the use of measurement-based peer supervision and comparing peer supervision with and without experts present.

METHODS

Setting. The study was nested in a randomized controlled trial conducted in North District of the state of Goa, India. The trial involved a sample of 280 pregnant women, aged ≥18 years with moderate to severe depressive symptoms, as defined by a PHQ-9 score≥10.

Participants. Three types of participants were included in the study:

Sakhis. Peers were mothers belonging to the same or neighbouring communities as the trial participants with depression. These were women with children who had, a similar socio-demographic background as participants, and good communication skills (Singla et al., 2014). These peers were recruited through *anganwadis* (village-based child development centers), advertisements in newspapers, or by word of mouth. Following the findings from our formative research,[15] selection criteria included being a mother (preferably with their youngest child over 3 years), an interest in helping other mothers in their community, good communication skills, a minimum seventh grade education and a commitment of one day per week. 55 peers were invited for interview, 37 peers were invited to training, and 26 peers entered and remained in the trial (see **Figure 1**). Here, competency assessment refers to Sakhis participation in standardized role plays to assess their skills following the attendance of a training workshop. Their mean age was 37.85 years (range 27 to 50 years), mean education was 11.85 years (range from 9 to 15 years) and they had about two children (range 1 to 3).

[INSERT FIGURE 1]

Experts. Experts included four mental health professionals with a Master's degree or diploma in clinical psychology with up to 5 years of experience. In total, there were six expert supervisors over the course of the trial and two at any given period. In addition, two lay counsellors with several years of experience in delivering psychological treatments for depression (Patel et al., 2017; Weobong et al., 2017) were recruited and trained as experts for the current study.

Participants. Study participants included pregnant women, aged ≥18 years with moderate to severe depressive symptoms, as defined by a Patient Health Questionnaire-9 (PHQ-9) score≥10 (Spitzer, Kroenke, Williams, & Group, 1999). Potentially eligible participants were screened for depression with a locally-validated version of the PHQ-9 (Patel et al., 2008) after providing written informed consent for screening (or witnessed informed consent/audio-recordings for illiterate participants). Participants were recruited from routine healthcare settings including two antenatal clinics and two primary health centers in Goa. Data from only those participants who were randomized to the intervention arm were selected for the current study. The trial protocols and results been described in full elsewhere (Fuhr et al., 2019; Sikander et al., 2015). The larger trial was registered on ClinicalTrials.gov (NCT02104232 in THPP-India).

Procedures

Thinking Healthy Program Peer-delivered. THPP is the adapted (peer-delivered) version of the Thinking Healthy Program (THP) which was originally developed and evaluated (based on delivery by government-employed LHWs) for perinatal women with moderate to severe depressive symptoms in Pakistan (Rahman et al., 2008).

THPP comprised up to 14 sessions over the perinatal period (antenatal and postnatal), each lasting up to 45 minutes. The intervention included four treatment phases: 1) the prenatal phase, in which the intervention was delivered during the second or third trimester of pregnancy in up to six sessions; 2) early infancy, in which the intervention was delivering during the first two months after childbirth in up to four sessions; middle infancy which occurred 3-4 months after childbirth over two sessions; and finally 4) late infancy, which occurred 5-6 months after childbirth over two sessions Treatment development and adaptation has been further described elsewhere (Rahman, 2007).

The core strategies used by the peers were both treatment-specific (behavioral activation, identifying, monitoring and replacing unhealthy behaviours with healthy ones and practicing them) as well as general (active listening, collaboration with the family, guided discovery using pictures and stories, homework; Atif et al., 2017). The treatment was implemented primarily in participants' homes and focused on three areas: personal well-being, relationship with the infant and relationship with significant others.

Training and supervision. Peers were trained in the THPP content, delivery, and therapy quality scales during the pilot phase of the trial (Aug – Dec 2013). Training included reviewing and practicing skills specific to antenatal and postnatal treatment phases as well as common treatment skills, such as taking a befriending stance, sharing one's own experiences as a mother, and building a good relationship with the mother and her family. The primary pedagogical tools were demonstrations, discussion and role plays. Following the training workshop, all Sakhis participated in a competency assessment whereby they engaged in a brief, ten to fifteen-minute standardized role play to demonstrate their skills. Independent raters evaluated Sakhis' performance on these role plays and Sakhis were selected for the trial based on their performance.

Upon completion of training and competency assessments, peers were assigned a caseload to practice the THPP programme and attended group supervision. Group supervision involved the assessment of therapy quality and followed the methods used in the PREMIUM trials (Singla et al., 2014) which were perceived as highly acceptable by lay health care workers. This involved listening to the individual audio-recorded counselling session in groups of 2 to 4 peers, followed by each peer rating the session using the Therapy Quality Scale (described below), and then discussing their ratings in detail among peers. During the pilot phase, supervision was led by the experts who had also trained peers on the THPP intervention.

During the trial itself, peer-led supervision was implemented with experts present in some sessions (but not leading the discussion). During the trial, two types of monthly group supervision (2-4 peers in each group) were encouraged: peer supervision with an expert present who helped to facilitate supervision and peer supervision without an expert present which were also referred to by the intervention team as 'mini meetings'; Golia & McGovern, 2013). During group supervision with experts, one audio-recorded session was independently rated by the provider who delivered the session, the peers, and the expert. Up to three-fourths of treatment sessions were audio-recorded; these audio recordings were randomly selected by an expert or the data coordinator and stratified by audio quality, phase and delivery agent. During group supervision without experts, Sakhis were placed into groups of 2-4 and asked to meet monthly to discuss treatment cases.

Measures and Data Collection.

The **Therapy Quality Scale** (TQS; Supplementary Figure 1) was developed and used for the assessment of individual sessions. The TQS is an 18-item scale that consists of two separate subscales: *Treatment Specific Skills* (TSS) measuring skills that are based on the THPP treatment modality (e.g., reviews previous session, assigns homework, involves family members); and *Treatment Approach Skills* (TAS) which assesses common factors that the counsellor may use in any psychotherapeutic intervention (e.g., using active listening, appropriate language and a collaborative approach). This scale was based on a 20-item therapy quality scale for the Healthy Activity Program (HAP), a culturally-adapted behavioral activation treatment delivered by lay health workers in primary healthcare centers in the study setting (Patel et al., 2017); it was found to have good psychometric properties of reliability (Singla et al., 2014). Adaptations to the SHARE TQS included treatment-specific skills related to THPP and each rating point being defined on the scale itself to enhance reliability by peers who would be expected to have less structured training and supervision. Each item is rated on a Likert scale (0 '*not done'*, 1 '*partially done'* or 2 '*done well'*). The score of each subscale was calculated as an average score of 2.

Patient Activation Scores. The PREMIUM Abbreviated Activation Scale (PAAS) is a 5-item scale, originally developed and used in the HAP trial (Weobong et al., 2017). It includes five indicators of behavioural activation — a treatment target in the THPP— such as the mother's self-report of her engagement with a variety of activities ("*did you engage in many different activities?*" and "*were you an active person and accomplished the goals you set out to do?*"). All five items are assessed on a scale of 0 (*'not at all'*) to 5 (*'yes, completely'*) for a total continuous score of 25. In the larger effectiveness trial, the PAAS at 3 months showed good internal consistency (α=0.801) and was found to mediate the effects of the THPP intervention on depressive symptoms (Singla et al., 2019).

Depression Outcome. The depression outcome used to estimate the predictive validity of therapy quality was depressive symptom severity scores on the Patient Health Questionnaire-9 (PHQ-9) at 6-months post-childbirth, as assessed by independent evaluators who were blind to allocation arm status. These data were recorded using tablets that were uploaded in real-time to a server with data being reviewed by independent data managers.

Qualitative Data. Two rounds of interviews at the beginning and end of the trial were conducted among all Sakhis and experts to examine their perspectives. Sakhis and experts were each interviewed twice by a pair of independent interviewers. Focus group discussion (FGDs) with Sakhis lasted up to 90 minutes, was conducted in September, 2015 in groups of 8 to 10. For the second round, in-depth interviews (IDIs) were conducted with Sakhis between January and July, 2017. Similarly, IDIs were used to collect data at the with experts lasted up to 45 minutes each, were conducted individually and took place in in January- February, 2016 and March-April, 2017.

Data Sources, Sample Sizes and Analyses.

Psychometric Properties of TQS. We examined the prosperities properties of the TQS by estimating its internal consistency, inter-rater reliability and test re-test reliability

Internal Consistency. Cronbach's alpha(α) was calculated to estimate the internal consistency across items of the TQS. All available ratings of facilitated peer group supervision sessions were used for these analyses, using the average peer rating of the same session (N=167).

Inter-rater reliability. Random pairs of expert ratings (n=44) for the same counselling session were selected. Because TQS is a continuous score, a one-way random effects analysis was used to calculate the intra-class correlation (ICC) as an estimate of inter-rater reliability between experts across the sessions (Cicchetti, 1994).

Test-retest reliability. Peers rated the same session twice (n=40): once at the time of the supervision, and re-rate the same audiotaped session between 7 to 30 days later[25]. The final sample size (N=40) provided 90% power (α=0.05) to detect a medium effect size on a Pearson correlation of r=0.30 (Cohen, 1988).

Predictive Validity. This analysis was restricted to mothers who have received at least one session in each of the antenatal and postnatal phases and who have completed their 3 months outcome assessment (estimated n=40 based on current enrollment and treatment drop-out rates). One randomly selected antenatal and one randomly selected postnatal counselling session tape will be rated by peers and the average rating used to derive a therapy quality rating for that mother. The analyses estimated the correlation between this average TQS rating and a) the PHQ-9 outcome score at 6 months, controlling for baseline severity of depressive symptoms; and b) the PAAS patient behaviour score at 3-months post-child birth. The final sample size (n=44) provided 90% power (α=0.05) to detect a small size on a Pearson correlation of r=0.30 (Cohen, 1988).

Reliability, Acceptability and Feasibility of Peer Supervision. Analyses related to reliability, acceptability and feasibility were conducted at mid-line (February 2016) and endline (October 2016) of the trial. This was because we wanted to use midline results to better understand and

potentially improve existing methods related to supervision. To estimate the reliability of peer ratings compared to expert ratings, the difference between the mean peer and expert TQS ratings was compared using a paired t-test. Because expert ratings were required, only sessions from facilitated peer supervision were used for these analyses.

Framework analysis (Ritchie, Spencer, & O'Connor, 2003) was used to analyze qualitative data, according to barriers and facilitators related to supervision processes from peers' and experts' perspectives. An iterative process was used to develop a coding tree. A team of five, independent research assistants charted the transcribed and translated data with high reliability (kappa=0.90) under supervision of two of the co-authors (DRS and RNK). Data was first charted separately by timepoint (midline or endline) and stakeholder group (Sakhi or expert), resulting in four charts. These data were triangulated to examine relevant subthemes across and within timepoints and stakeholder groups. Data was then charted according to the major themes of barriers and facilitators under the wider category of supervision followed by the development of subthemes within the general category of barriers and facilitators (CR and DRS). A final review of the qualitative coding was conducted by two of the co-authors (DRS and CR).

RESULTS

Psychometric properties of the Therapy Quality Scale

Overall, the TQS demonstrated excellent reliability and validity. The internal consistency of the scale across all items was psychometrically sound for both treatment-specific skills (α=0.801) and general skills (α=0.816). ICC estimates between expert ratings of the same session showed moderate inter-rater reliability on treatment-specific skills (ICC=0.707) and excellent inter-rater reliability on general skills (ICC=0.904). Test-retest reliability of peer ratings on the same session demonstrated a moderate correlation between time-points on both treatment-specific skills (r=0.484, p<0.01) and general skills (r=0.501, p<0.001). Finally, the TQS was significantly related to patient activation scores at 3-months post-child birth for both treatment-specific skills (r=0.375, p<0.01) and general skills (r=0.313, p<0.01); the relations between TQS and depressive symptoms scores was not significant. Patient activation scores were also significantly related to patient depressive symptom scores at 6-months post-child birth (r=-0.455, p<0.001).

Accuracy of Peer Ratings vs. Expert Ratings

In total, 167 ratings were used to assess the accuracy of peer ratings compared to expert ratings during facilitated peer supervision sessions (n=92 by midline and n=75 by endline). Across 26 Sakhis, 7.13 sessions (range 2.6 to 7.7) were rated for each counsellor; session duration lasted an average of 38.48 minutes (95% CI=36.50 to 40.46 minutes) and half of the sessions (n=86 or 51.5%) were derived from the prenatal phase of treatment.

Mid-line assessments (February 2016) demonstrated significant differences between average peer and expert ratings on both treatment-specific and therapeutic skills—demonstrating that peers—on average—were not assessing individual audio sessions as reliably as experts (see Figure 2a). This resulted in several, key modifications in supervision practices including the introduction of refresher trainings (which involved extra group training by an expert on the TQS to clarify each item), stopping and discussing audio-recording in chunks

rather than listening to the whole tape before conducting the rating; and the examination of Sakhis' perspectives on specific items and their relationship to the overall THPP program. In addition, two of the expert supervisors who previously served as lay counsellors in the HAP trial shared their experiences and strategies in graduating from a lay person to a counsellor and used the language of the Sakhis to ensure that they understood the core THPP constructs of activation and peer support. Finally, individual feedback was provided to those Sakhis whose scores were lower than the peer average for that item.

At the end of the trial (October 2017), the accuracy of peer ratings compared to experts was again assessed. Paired t-tests demonstrated no significant difference on either treatment-specific or general skills (Figure 2b), indicating that Sakhis had learned to accurately rate sessions as reliably as experts.

[INSERT FIGURE 2A AND 2B]

Acceptability and Feasibility of Peer Supervision among Peer Delivery Agents

Qualitative findings were largely consistent between Sakhis and experts (see Table 1) with some discrepancies. Perceived barriers and facilitators of peer supervision from both of these perspectives and across the trial are discussed below.

Barriers. Initially, the majority of both Sakhis and supervisors reported difficulties with travel due to a lack of compensation and a lack of time to attend meetings due to other commitments, especially in supervision where experts where not present. These barriers remained the most endorsed across the duration of the trial.

"Mini Meetings [supervision without experts] are time consuming and also travel allowances aren't paid to us and neither are expenses… We visit once in a month and plus we conduct mini supervision ourselves so this is not productive for us." (Sakhi_23, Round 1)

"Sakhis say that they don't have time—that is why they don't have [mini-meetings]. But we encourage them to have these [mini-meetings] as they will benefit from it." (Supervisor_03, Round 1)

Both experts and Sakhis reported that supervision without experts often lacked relevant feedback, and that there was overall low motivation to attend these sessions. This remained unchanged throughout the trial despite attempting to address challenges such as travel allowances.

"In absence of a supervisor, there is no control in our discussion also and we don't rate the session. So I don't feel it benefits us but consumes time." (Sakhi_20, Round 1)

"When there are difficulties that have been brought up again there was a chance that the peer would give wrong feedback. Because there would be a wrong answer which couldn't be corrected because we [Supervisors] weren't there." (Supervisor_02, Round 2)

Other key barriers reported by both stakeholder groups reflected peers' initial difficulties in rating the TQS and leading supervision as well as being score a '0' to score individual items on the TQS. Initially, Sakhis reported feeling demotivated when rated '0' on an item. By endline, a number of barriers such as initial difficulties with TQS items, a lack of financial incentive, competing demands and a low quality of feedback were less frequently mentioned.

"At first, Sakhis didn't understand the rating scale. But then, we told them to read each item and then we would discuss. […] After reading, they understand the meaning. But at first, they would fill in [the rating scale] because they have to do it without understanding." *(Supervisor_01, Round 1)*

"[…] After listening [to my recording] if all say that it was good then I feel very nice. If they rate me 0, then I feel bad thinking that I worked hard and still I got rated 0." (Sakhi_17, Round 1)

Facilitators. The most commonly reported facilitators by Sakhis were the benefit of using a structured scale to rate therapy quality that in turn facilitated consistent and relevant feedback and fostered a greater understanding of Sakhis' challenges and skills. This was also endorsed by all supervisors and across the duration of the trial.

"[Structured supervision] has helped a lot. We understand and learn from our mistakes. We get to know how we can deal with the problems…so there is improvement in our sessions." (Sakhi_20, Round 2)

"When we fill in the TQS, we come to know where Sakhis are facing challenges and where they have done a good job. Suppose they have difficulties, we come to know and where else they need more training. That is why I feel it is very important." (Supervisor_03, Round 1)

A major facilitator reported by both Sakhis and experts in supervision with experts was also the positive group processes, where both expressed that this was a chance to meet and solve common problems before bringing it up with the supervisor.

"When all the Sakhis would meet [for group supervision] there used to be good discussions on how to handle the difficult cases or any case. We would get an idea about it." (Sakhi_11, Round 2)

"Group supervision is better since there are just 2-4 in a group. It is better because we can ask them what difficulties they face. […] And if they cannot do something, then they ask us. Since they are few in numbers, they ask us in what better way they can do it [the session] well." (Supervisor_02, Round 1)

In general, both experts and Sakhis preferred the expert to be present to ensure that Sakhis attend supervision, that the information being shared was accurate, and to help with difficult cases. Similarly, both experts and Sakhis noted that living close to one another was a primary facilitator for peer supervision without an expert present.

"When they [Sakhis] live close by, then mini-meetings happen very nicely. They make time and sit together or those who are at home only don't go for work. Their mini-meetings take place because the Sakhis are at home and don't go for work." (Supervisor_04, Round 1)

"…currently our group is really nice because we live close to each other so it is convenient for meetings, so we do mini meeting at my home or theirs." (Sakhi_22, Round 2)

DISCUSSION

This study describes our efforts to systematically examine the accuracy and acceptability of peer supervision among peer delivery agents within a randomized controlled trial for perinatal depression in Goa, India. Specifically, we developed and estimated the psychometric properties of the SHARE Therapy Quality Scale (TQS); we examined the extent to which peers could be trained to evaluate their peer counselling skills as reliably as experts; and we conducted a qualitative study to examine relevant barriers and facilitators of peer supervision with and without experts.

Psychometric Properties of Therapy Quality Scale

We found robust and consistent evidence for the TQS to be psychometric reliable and valid. Specifically, we found evidence for good internal consistency, test-retest and inter-rater reliability of the TQS. In addition, we found good predictive validity between the TQS and subsequent patient behaviours at 3-months post-child birth but not on clinical outcomes at 6-months post-childbirth. Our findings confirm that therapy quality aligns with a temporal pathway: higher therapy quality was related to better patient behaviours (in this case, improved patient activation scores) which in turn were related to better clinical outcomes (lower depressive symptoms). This temporal pathway has been demonstrated in other psychological treatments for depression and alcohol treatment trials in both LMICs (Singla et al., 2019) as well as high-income countries (HICs; DeRubeis et al., 1990; Magill et al., 2014).

Rating Sessions Reliably

With time, we found that peers were eventually able to rate their audio-recorded sessions as reliably as expert supervisors. Our findings confirmed our primary hypothesis of increased levels of agreement between peer and expert ratings as demonstrated by a decrease in the difference in mean therapy quality scores between raters to non-statistical differences. This finding was common in both the assessment of general skills as well as treatment-specific skills, illustrating peers' consistent assessment in both types of skills.

However, it is important to note that similar to gaining ~~competency~~ an aptitude to ~~in~~ deliver~~ing~~ treatment sessions, it took time for peers to be able to rate sessions as reliably as experts. As indicated by our results at midline, this did not occur immediately and additional training was required within the larger trial to help foster Sakhis' ability to better understand and utilize this tool in order to accurately rate sessions. In previous studies. This is similar to other studies which have tested the same question which involved a three-month internship phase (Singla et al., 2014). In short, these findings suggest that experts are required to facilitate peer supervision until reliability is achieved.

Our qualitative study highlighted that this model of measurement-based supervision was found to be acceptable: It promoted accurate feedback between peers and from experts, utilized a group process to facilitate a productive and supportive discussion and was found to be helpful in skill development particularly when experts were present. These findings speak to the general psychological treatment literature which has highlighted the benefits of a shared developmental process and vicarious learning as both observers and learners within group peer supervision (Borders, 2012), as well as the development of increased self-awareness and skill building (Wheeler & Richards, 2007). They also replicate our findings from a lay-counsellor delivered intervention for depression which highlighted the preference for structured feedback (Singla et al., 2014).

Preference for Supervision with Experts

In addition, peer supervision without experts was reported to be less feasible than supervision with experts (also referred to as planned and facilitated supervision, respectively; Golia & McGovern, 2013). Despite a preference for supervision with experts—both among experts and Sakhis—it is important to note that geographical proximity was considered a key facilitator for peer supervision without experts. Additional studies are required to examine how to foster the systematic and scaled use of measurement-based supervision to facilitate peer supervision without having to rely on the presence of expert supervisors. For example, the use and testing of digital platforms may be helpful to overcome such structural barriers in facilitating peer supervision (Naslund, Shidhaye, & Patel, 2019). ~~For example,~~Recent evidence of digital training platforms was assessed in a recent non-inferiority trial in Pakistan where ~~demonstrated that~~ competency scores among community health workers ~~in~~ delivering the Thinking Healthy Programme ~~by community health workers~~ were equivalent among those trained in the digital training arm when compared to face-to-face training (Rahman et al., 2019).

Strengths. To the best of our knowledge, this is among the first systematic efforts to examine and replicate methods to examine the quality with which non-specialists deliver psychological treatments. We utilized methods that were previously tested in the Healthy Activity Program (HAP) which examined lay counsellors (Singla et al., 2014)—and extended those findings to include peers and examining peer supervision models without experts present. For example, in a recent review of 22 studies examining supervision methods for community health workers in low- and middle-income countries, none assessed the quality with which community health workers (CHWs) delivered their programs (Hill et al., 2014); this finding is consistent in high-income countries as well (Borders, 2012). Additional strengths of our study include the rigorous examination of psychometric properties of the TQS which was found to be reliable and valid. Furthermore, our qualitative study examined the relevant barriers and facilitators of peer group supervision, highlighting the preference for supervision with experts present.

Limitations. Despite the strength of our study, we acknowledge several limitations. First, peers and experts were not blind to the identity of the peer whose sessions was being rated. Second, the therapy quality ratings of peer supervision without experts were not collected, and therefore cannot be compared quantitatively with peer supervision with experts present.

Finally, despite the important findings highlighting the accuracy and acceptability of peer-based supervision models, our methods require additional replication of the TQS scale among peer and other non-specialist delivered treatments as well as future examination of the use of digital platforms to enhance the acceptability and scalability of measurement-based, peer supervision.

In conclusion, peer delivery agents can be trained to implement peer supervision and, with time, assess therapy quality as reliably as experts using a psychometrically-sound and acceptable measure of assessment. Despite these achievements, our results also highlight that the presence of experts is preferred as a facilitator of peer-delivered supervision. In order to facilitate peer supervision without experts, addressing structural challenges including distances between Sakhis, costs and time related constraints and ensuring adequate feedback between Sakhis are needed. Additional studies are required to examine these findings to ultimately contribute to the scalability of non-specialist delivered mental health care globally.

Acknowledgements

We thank all of the participants, treatment providers and expert supervisors who participated in this study. In addition, we would like to thank Anisha Lazarus, Ejma Afonso, Sherline Dias, Shamin Colaco, Tulsi Shet, Christalina Monteiro, Kirti Gawas, Ankita Volvoikar for their efforts in conducting interviews and transcribing the qualitative data as well as Ethel D'Souza for organizing the therapy quality ratings. The research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under award number 1U19MH095687-S1. Daisy R. Singla is currently supported by an Academic Scholars Award from the Department of Psychiatry at the University of Toronto.

References

Atif, N., Krishna, R. N., Sikander, S., Lazarus, A., Nisar, A., Ahmad, I., . . . Rahman, A. (2017). Mother-to-mother therapy in India and Pakistan: adaptation and feasibility evaluation of the peer-delivered Thinking Healthy Programme. *BMC psychiatry, 17*(1), 79.

Borders, L. D. (2012). Dyadic, triadic, and group models of peer supervision/consultation: What are their components, and is there evidence of their effectiveness? *Clinical Psychologist, 16*(2), 59-71.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment, 6*(4), 284.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences 2nd edn. In: Erlbaum Associates, Hillsdale.

DeRubeis, R. J., Evans, M. D., Hollon, S. D., Garvey, M. J., Grove, W. M., & Tuason, V. B. (1990). How does cognitive therapy work? Cognitive change and symptom change in cognitive therapy and pharmacotherapy for depression. *Journal of Consulting and Clinical Psychology, 58*(6), 862.

Fairburn, C. G., & Cooper, Z. (2011). Therapist competence, therapy quality, and therapist training. *Behaviour research and therapy, 49*(6), 373-378.

Fairburn, C. G., & Patel, V. (2017). The impact of digital technology on psychological treatments and their dissemination. *Behaviour research and therapy, 88*, 19-25.

Fuhr, D. C., Weobong, B., Lazarus, A., Vanobberghen, F., Weiss, H. A., Singla, D. R., . . . Patel, V. (2019). Delivering the Thinking Healthy Programme for perinatal depression through peers: an individually- randomised controlled trial in India *The Lancet Psychiatry, 6*(2), 115-127.

Golia, G. M., & McGovern, A. R. (2013). If you save me, I'll save you: the power of peer supervision in clinical training and professional development. *The British Journal of Social Work, 45*(2), 634-650.

Herschell, A. D., Kolko, D. J., Baumann, B. L., & Davis, A. C. (2010). The role of therapist training in the implementation of psychosocial treatments: A review and critique with recommendations. *Clinical Psychology Review, 30*(4), 448-466.

Hill, Z., Dumbaugh, M., Benton, L., Kaellander, K., Strachan, D., Ten Asbroek, A., . . . Meek, S. (2014). Supervising community health workers in low-income countries–a review of impact and implementation issues. *Global health action, 7*(1), 24085.

Hoeft, T. J., Fortney, J. C., Patel, V., & Unützer, J. (2016). Task-Sharing Approaches to Improve Mental Health Care in Rural and Other Low-Resource Settings: A Systematic Review. *The Journal of rural health*.

Kazdin, A. E. (2017). Addressing the treatment gap: A key challenge for extending evidence-based psychosocial interventions. *Behaviour research and therapy, 88*, 7-18.

Ladany, N., Ellis, M. V., & Friedlander, M. L. (1999). The supervisory working alliance, trainee self-efficacy, and satisfaction. *Journal of Counseling & Development, 77*(4), 447-455.

Magill, M., Gaume, J., Apodaca, T. R., Walthers, J., Mastroleo, N. R., Borsari, B., & Longabaugh, R. (2014). The technical hypothesis of motivational interviewing: A meta-analysis of MI's key causal model. In: American Psychological Association.

Naslund, J. A., Shidhaye, R., & Patel, V. (2019). Digital Technology for Building Capacity of Nonspecialist Health Workers for Task Sharing and Scaling Up Mental Health Care Globally. *Harvard review of psychiatry, 27*(3), 181-192.

Patel, V., Araya, R., Chowdhary, N., King, M., Kirkwood, B., Nayak, S., . . . Weiss, H. A. (2008). Detecting common mental disorders in primary care in India: a comparison of five screening questionnaires. *Psychological medicine, 38*(02), 221-228.

Patel, V., Weobong, B., Weiss, H. A., Anand, A., Bhat, B., Katti, B., . . . King, M. (2017). The Healthy Activity Program (HAP), a lay counsellor-delivered brief psychological treatment for severe depression, in primary care in India: a randomised controlled trial. *The Lancet, 389*(10065), 176-185.

Rahman, A. (2007). Challenges and opportunities in developing a psychological intervention for perinatal depression in rural Pakistan–a multi-method study. *Archives of women's mental health, 10*(5), 211-219.

Rahman, A., Akhtar, P., Hamdani, S. U., Atif, N., Nazir, H., Uddin, I., . . . Sikander, S. (2019). Using technology to scale-up training and supervision of community health workers in the psychosocial management of perinatal depression: a non-inferiority, randomized controlled trial. *Global Mental Health, 6*.

Rahman, A., Malik, A., Sikander, S., Roberts, C., & Creed, F. (2008). Cognitive behaviour therapy-based intervention by community health workers for mothers with depression and their infants in rural Pakistan: a cluster-randomised controlled trial. *The Lancet, 372*(9642), 902-909.

Ritchie, J., Spencer, L., & O'Connor, W. (2003). Carrying out qualitative analysis. *Qualitative research practice: A guide for social science students and researchers, 2003*, 219-262.

Saxon, D., Barkham, M., Foster, A., & Parry, G. (2016). The Contribution of Therapist Effects to Patient Dropout and Deterioration in the Psychological Therapies. *Clinical Psychology & Psychotherapy*.

Saxon, D., Firth, N., & Barkham, M. (2016). The relationship between therapist effects and therapy delivery factors: Therapy modality, dosage, and non-completion. *Administration and Policy in Mental Health and Mental Health Services Research*, 1-11.

Sikander, S., Ahmad, I., Atif, N., Zaidi, A., Vanobberghen, F., Weiss, H. A., . . . Rahman, A. (accepted). Delivering the Thinking Healthy Programme for perinatal depression through volunteer peers: a cluster randomised controlled trial in Pakistan. *The Lancet Psychiatry*.

Sikander, S., Lazarus, A., Bangash, O., Fuhr, D. C., Weobong, B., Krishna, R. N., . . . Rahman, A. (2015). The effectiveness and cost-effectiveness of the peer-delivered Thinking Healthy Programme for perinatal depression in Pakistan and India: the SHARE study protocol for randomised controlled trials. *Trials, 16*(1), 534.

Singla, D., Lazarus, A., Atif, N., Sikander, S., Bhatia, U., Ahmad, I., . . . Patel, V. (2014). "Someone like us": Delivering maternal mental health through peers in two South Asian contexts. *Journal of affective disorders, 168*, 452-458.

Singla, D. R., Hollon, S. D., Velleman, R., Weobong, B., Nadkarni, A., Fairburn, C. G., . . . Patel, V. (2019). Temporal pathways of change in two randomized controlled trials for depression and harmful drinking in Goa, India. *Psychological medicine*, 1-9. doi:10.1017/S0033291718003963

Singla, D. R., Kohrt, B., Murray, L. K., Anand, A., Chorpita, B. F., & Patel, V. (2017). Psychological treatments for the world: lessons from low-and middle-income countries. *Annual review of clinical psychology, 13*(1).

Singla, D. R., Kohrt, B. A., Murray, L. K., Anand, A., Chorpita, B. F., & Patel, V. (2017). Psychological Treatments for the World: Lessons from Low- and Middle-Income Countries. *Annu Rev Clin Psychol, 13*, 149-181. doi:10.1146/annurev-clinpsy-032816-045217

Singla, D. R., & Kumbakumba, E. (2015). The development and implementation of a theory-informed, integrated mother-child intervention in rural Uganda. *Social Science & Medicine, 147*, 242-251.

Singla, D. R., MacKinnon, D. P., Fuhr, D. C., Sikander, S., Rahman, A., & Patel, V. (2019). A multiple mediation analysis of the peer delivered Thinking Healthy Program for perinatal depression: Findings from two parallel randomized controlled trials. *British Journal of Psychiatry*, 1-8.

Singla, D. R., 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 peer-led therapy quality assessment in Goa, India. *Behaviour research and therapy, 60*, 53-59.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Group, P. H. Q. P. C. S. (1999). Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *Jama, 282*(18), 1737-1744.

Van Ginneken, N., Tharyan, P., Lewin, S., Rao, G. N., 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 Library*.

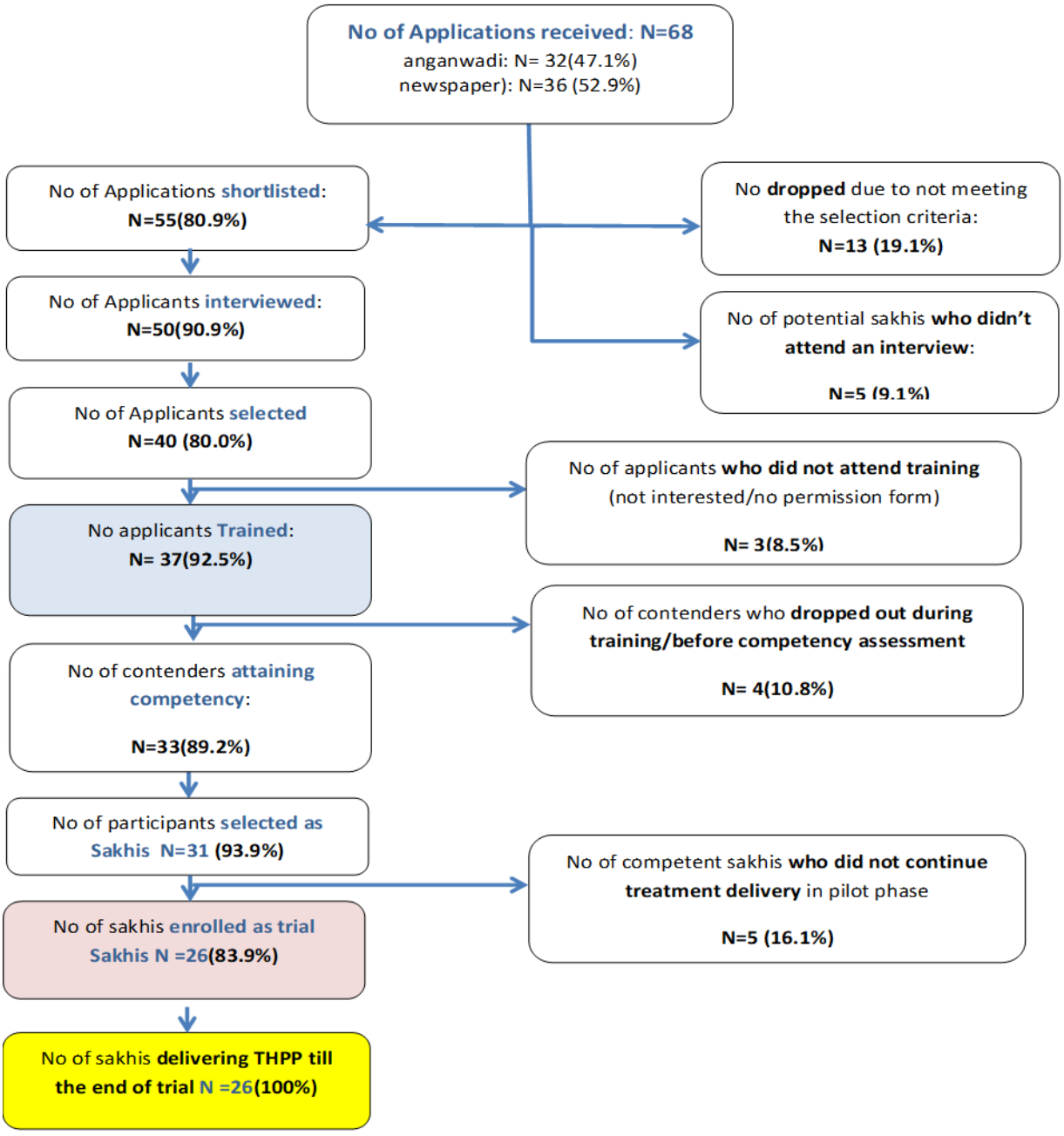
Watkins Jr, C. E. (2011). Does psychotherapy supervision contribute to patient outcomes? Considering thirty years of research. *The Clinical Supervisor, 30*(2), 235-256.

Weobong, B., Weiss, H. A., McDaid, D., Singla, D. R., Hollon, S. D., Nadkarni, A., . . . Patel, V. (2017). Sustained effectiveness and cost-effectiveness of the Healthy Activity Program, a brief psychological treatment for depression delivered by lay counsellors in primary care: twelve-month follow-up of a randomised controlled trial. *PLoS Medicine, 17*(9).

Wheeler, S., & Richards, K. (2007). The impact of clinical supervision on counsellors and therapists, their practice and their clients. A systematic review of the literature. *Counselling and Psychotherapy Research, 7*(1), 54-65.

Yeh, C. J., Chang, T., Chiang, L., Drost, C. M., Spelliscy, D., Carter, R. T., & Chang, Y. (2008). Development, content, process and outcome of an online peer supervision group for counselor trainees. *Computers in Human Behavior, 24*(6), 2889-2903.

662 **Figure 1. Recruitment, Training and Retention of Peers.**
663



664
665

Figure 2a. Accuracy of Peer Ratings of Audio-Recorded Treatment Sessions at Midline (February 28 2016; N=92)

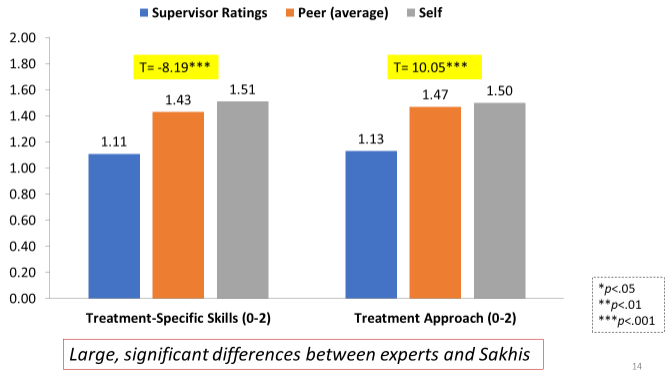
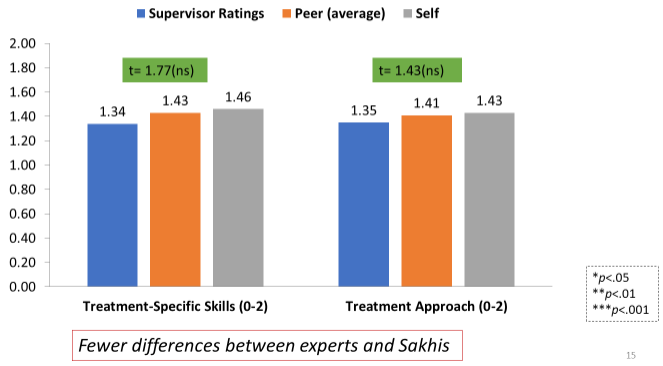


Figure 2b. Accuracy of Peer Ratings of Audio-Recorded Treatment Sessions at Endline (October 15, 2016; N=75)



673 **Table 1. Qualitative findings from the perspectives of Sakhis and Experts.**
674

	Sakhis		Experts	
	Round 1	Round 2	Round 1	Round 2
Supervision				
Barriers				
Geographical and travel barriers	+++	+	+++	+
Initial difficulties with specific items in TQS	+++	-	+++	+
Impact of "Zero" Rating	++	+	++	+
Low motivation to attend without experts	+++	+++	++	++
Competing Demands/No Time	++	+	++	-
No financial incentives	++	++	+	+
Facilitators				
TQS - Structured Feedback and Increased Awareness of Challenges/Skills	+++	+++	+++	++
Group Processes	+++	+++	+++	++
Preference for Supervision with Experts	++	+++	+	+++
Reliance on Experts for Support	+	++	-	-

675
676 "-" = no endorsement
677 "+" = some endorsement (<25%)
678 "++" = good endorsement (25-59%)
679 "+++" = large majority endorsement (>60%)
680
681