Measurement and assessment of fidelity and competence in nonspecialist-delivered, evidence-based behavioral and mental health interventions: A systematic review

Laura Bond [*], Erik Simmons, Erika L. Sabbath

Boston College School of Social Work, 140 Commonwealth Avenue, Chestnut Hill, MA, 02467, USA

ABSTRACT

Nonspecialists have increasingly been used to deliver evidence-based, mental health and behavioral interventions in lower resource settings where there is a dearth of specialized providers and a corresponding gap in service delivery. Recent literature acknowledges that nonspecialist-delivered interventions are shown to be effective. However, few studies report on the fidelity (the degree to which an intervention was implemented as intended) and/or competence (general skills of nonspecialists), key concepts that measure quality of evidence-based intervention delivery. This study seeks to understand how both fidelity and competence have been assessed in nonspecialist-delivered, evidence-based interventions with an intended social or psychological behavior-change outcome. Our search results originally yielded 2317 studies, and ultimately, 16 were included in our final analysis. Generally, results from a narrative synthesis indicated that tools used in the studies demonstrated sufficient inter-rater reliability and intra-class correlation components. Included studies used and described a range of fidelity and competence tools. However, the ENhancing Assessment of Common Therapeutic factors tool was the most commonly used tool that measures competence of nonspecialists, and has been adapted to several other settings. The roles of supervisors in mentoring, monitoring, and supervising nonspecialists emerged as a key ingredient for ensuring fidelity. Most studies assessing fidelity were limited by small sample sizes due to low numbers of nonspecialists implementing interventions, however, more advanced statistical methods may not be needed and may actually impede community-based organizations from assessing fidelity data. Our results suggest interventions can share resources, tools, and compare findings regardless with proper supervision. While the two terms "fidelity" and "competence" are often used interchangeably, their differences are noteworthy. Ultimately, both competency and fidelity are critical for delivering evidence-based interventions, and nonspecialists are most effective when they can be evaluated and mentored on both throughout the course of the intervention.

1. Introduction

Nearly 10% of the global population faces a mental health disorder at any point in time, and yet only 1% of the global health workforce is equipped to provide care for mental and behavioral health challenges (Keynejad et al., 2018). The World Health Organization (WHO) has responded to the gap in mental health service delivery by launching the Mental Health Gap Action Programme (mhGAP), which provides evidence-based guidance for delivering and scaling up mental health

interventions, and acknowledges a growing body of evidence that mental health interventions can be delivered by trained and supervised nonspecialists alike (World Health Organization & UN High Commissioner for Refugees, 2015). In addition, projects such as Ensuring Quality in Psychological Support (EQUIP) and digital platforms like EMPOWER have been designed by lead researchers and stakeholders in the fields of global health to consider ways to scale out tools and training resources for evidence-based, behavioral interventions that are delivered by nonspecialists (The President and Fellows of Harvard College,

2022; World Health Organization, 2022). These digital platforms provide online curriculum and toolkits for evidence-based, mental health interventions (The President and Fellows of Harvard College, 2022; World Health Organization, 2022).

Given a dearth of specialized providers in many settings, nonspecialists are critical for delivering evidence-based interventions (EBIs). Task-shifting refers to specialists collaborating with nonspecialist providers to deliver health-related services that have traditionally been assigned to experts with professional training and certification (WHO, 2007; World Health Organization, 2008). Task-shifting in the field of global mental health has been an effective and increasingly prevalent strategy that addresses the shortage of mental health specialists in low and middle-income countries (LMICs) (Tsai, 2017), particularly as multi-sectoral approaches have demanded more comprehensive health systems which involve nonspecialists (Kakuma et al., 2014; Leocata et al., 2021).

In the past two decades, non-specialized and non-formally trained individuals (sometimes referred to also as lay workers or community health workers) have successfully delivered a range of mental health and behavioral interventions, including early childhood development and family violence reduction home-visiting programs (Barnart et al., 2020; Desrosiers et al., 2021), interpersonal psychotherapy for depression and anxiety disorders (Betancourt et al., 2021; Bolton et al., 2003; Newnham et al., 2015; Patel et al., 2010, 2011, 2017), and alcohol use disorder treatments (Nadkarni et al., 2017; Sileo et al., 2021). A recent systematic review revealed that nonspecialist-delivered interventions have demonstrated improvements in mental health and behavioral outcomes with moderate to large effect sizes (Singla et al., 2017).

Although nonspecialist-delivered interventions are acknowledged as common strategies that effectively bridge the treatment gap and reduce health disparities, particularly in lower-resource settings, less attention has been paid to the quality of training, supervision, and fidelity and/or competence to evidence-based treatment programs in these settings (Shahmalak et al., 2019; Kanzler et al., 2021; Singla et al., 2017; Kohrt et al., 2015). A 2015 systematic review which identified quantitative instruments of implementation outcomes (acceptability, adoption, appropriateness, cost, feasibility, fidelity, penetration, sustainability) within mental or behavioral health settings found no fidelity instruments which included "either assessments of implementation interventions (e.g., instruments that measure frequency and structure of an evidence-based practice training) or instruments that could be applied to any evidence-based practice" (Lewis et al., 2015, p. 8). Thus, researchers may be ill-equipped with resources to systematically and reliably measure fidelity in their studies.

Often, studies that described the process of training and supervision of nonspecialists have not discussed if or how these efforts resulted in fidelity and/or competence to the evidence-based intervention, nor are there standardized tools or measures for these constructs that can be shared across interventions (Ginsburg et al., 2021). Indeed, attention to treatment fidelity or therapist competence increases the burden on researchers and agencies, requiring greater investments in time, equipment, and personnel. Nonetheless, O'Shea et al. argue that assessment of treatment fidelity is cost-effective in the long-term, leading to higher-quality, reliable care, and ensuring an efficient translation of evidence-based practices into routine care (2016). As the use of nonspecialists continues, it will be critical to address issues of fidelity, competence, and ultimately, quality, in order to expand access to equitable health treatment.

1.1. Key concepts

Evidence-Based, Behavioral Interventions. Evidence-based interventions are interventions that have an established causal relationship between the intervention outputs and the intended outcomes in the population and delivery setting (Leeman et al., 2017). Leeman and colleagues define evidence-based interventions as "any action or set of

actions that delivery systems enact to improve health behaviors, health outcomes, or health-related environments (e.g., built and communication environments that support healthy behaviors)" (p. 3). Behavior-change interventions are a subset of evidence-based interventions, and are defined as "coordinated sets of activities designed to change specified behavior patterns" (Michie et al., 2011). Behavioral interventions can be used to promote uptake of healthy lifestyles or practices, address ongoing mental health challenges and provide relevant coping strategies, or promote family strengthening or positive parenting practices.

Implementation Fidelity. Implementation fidelity was first identified as a critical issue, as scholars noted distance between the intended purpose of a program or policy, and its implementation (Elmore, 1980; Crea et al., 2009). Later, several conceptual distinctions for implementation outcomes emerged. Proctor et al. defined the concept of fidelity as "the degree to which an intervention was implemented as it was prescribed in the original protocol or as it was intended by the program developers" (Proctor et al., 2011). Resnick et al. (2005) also contributes to an operationalization of fidelity, defining fidelity as the "methodological strategies used to monitor and enhance the reliability and validity of behavioral interventions." In describing the defining characteristics of implementation research applied to global health settings, Theobald et al. (2018) explain how a focus on processes and outcomes allows implementation researchers to engage stakeholders and to assess fidelity, among other implementation outcomes. They define fidelity as "implementation according to its (the evidence-based intervention) design" (p. 2225). The term "adherence" is often used as a synonym for fidelity, and also refers to "the extent to which a therapist used interventions and approaches prescribed by the treatment manual" (Waltz et al., 1993, p. 620). Carroll and colleagues (2007) suggest that adherence is a part of fidelity, not a synonym, and that fidelity also consists of other outcomes that relate to overall quality of delivery, such as participant responsiveness and exposure or dosage.

Implementation Competence. Competence speaks to the general skills of nonspecialist-facilitated interventions rather than intervention-specific skills (Kohrt et al., 2015). The overall quality of intervention delivery is dependent upon both fidelity and competence (Fairburn & Cooper, 2011). It is critical to distinguish competence from fidelity, as the terms are used interchangeably and studies examining fidelity often examine competence instead of, and in addition to, fidelity (Ottman, Kohrt, Pedersen, & Schafer, 2020). Both fidelity and competence are critical for therapists when delivering an intervention, and distinct enough to be measured separately when possible. Many scholars define competence as the "common factors" that all therapies have in common (Cuijpers et al., 2019; Wampold, 2015; Kohrt et al., 2015). The term "global competence" emerges as the idea that a broad range of soft skills can be harnessed by facilitators to manage problems and assist intervention participants with realizing their goals (Barber, Sharpless, Klostermann, & McCarthy, 2007; Ottman et al., 2020). These skills are not intervention-specific, but relevant to all mental health and psychosocial support interventions. Competencies may include skills like showing empathy, active listening, or adapting an activity to better meet a participants' needs.

Nonspecialists. A number of terms have been used in the literature to describe nonspecialists delivering an intervention, including but not limited to "layworkers," "paraprofessionals," "peer counselors," "community health workers," "lay counselors," "village health workers," "health promotores," and "auxiliary health staff" (Lehmann & Sanders, 2007; Kanzler et al., 2021, p. 4). The World Health Organization defines nonspecialists as anyone who "was trained in some way in the context of the intervention; but has received no formal professional or paraprofessional certificate or tertiary education degree" (World Health Organization, 2007, p. 79).

1.2. Objectives

This study adheres to Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) (Moher et al., 2015). In this systematic review, we seek to understand how fidelity and/or competence has been assessed in nonspecialist-delivered, evidence-based interventions, illuminating the relationship between fidelity and competence and how this relationship manifests in the tools intended to measure quality of delivery. Specifically, our review will focus on identifying tools and methods that have been used to collect fidelity and/or competence data and understanding how the concepts have been measured in evidence-based, psychosocial and behavioral interventions. In such interventions, we will look for who is responsible for monitoring fidelity and/or competence and how often, unique or shared characteristics of the monitoring tools used, the setting that data is collected in, and which methods teams are using to analyze fidelity data after it has been collected.

2. Methods

2.1. Search strategy and inclusion criteria

This study followed PRISMA guidelines (see Fig. 1 for a PRISMA flowchart of the screening process and results). After registering the study on Open Science Framework in October 2021, authors performed key word searches in October 2021 in PubMed, APA PsycInfo, Sociological Abstracts, and SCOPUS databases (see Table 1 for key words used in each of the searches). Eligibility criteria for studies included: a) interventions delivered by a nonspecialist, b) studies using evidence-based interventions with an intended social or psychological behavioral

Table 1
Search strategy.

| Key Word	Search Terms
Nonspecialist	(nonprofessional* OR "non specialist*" OR nonspecialist* OR "para professional*" OR paraprofessional* OR (peer (coach OR counselor OR counselor OR educat* OR facilitat*)) OR (lay (counselor OR counselor OR "health worker")) OR (community ("health worker" OR facilitator OR organi*)))
Evidence-based behavioral intervention	(Intervention* OR treatment* OR program* OR "evidence based" OR "behav* intervention", "behav* change intervention" OR ("Psychotherapy") OR ("Behavior Modification") OR ("Evidence Based Practice") OR ("Program Evaluation"))
Fidelity or Competence	((fidelity OR competence OR adherence OR reliability) AND (tool OR instrument OR assessment OR analysis OR measure OR checklist))
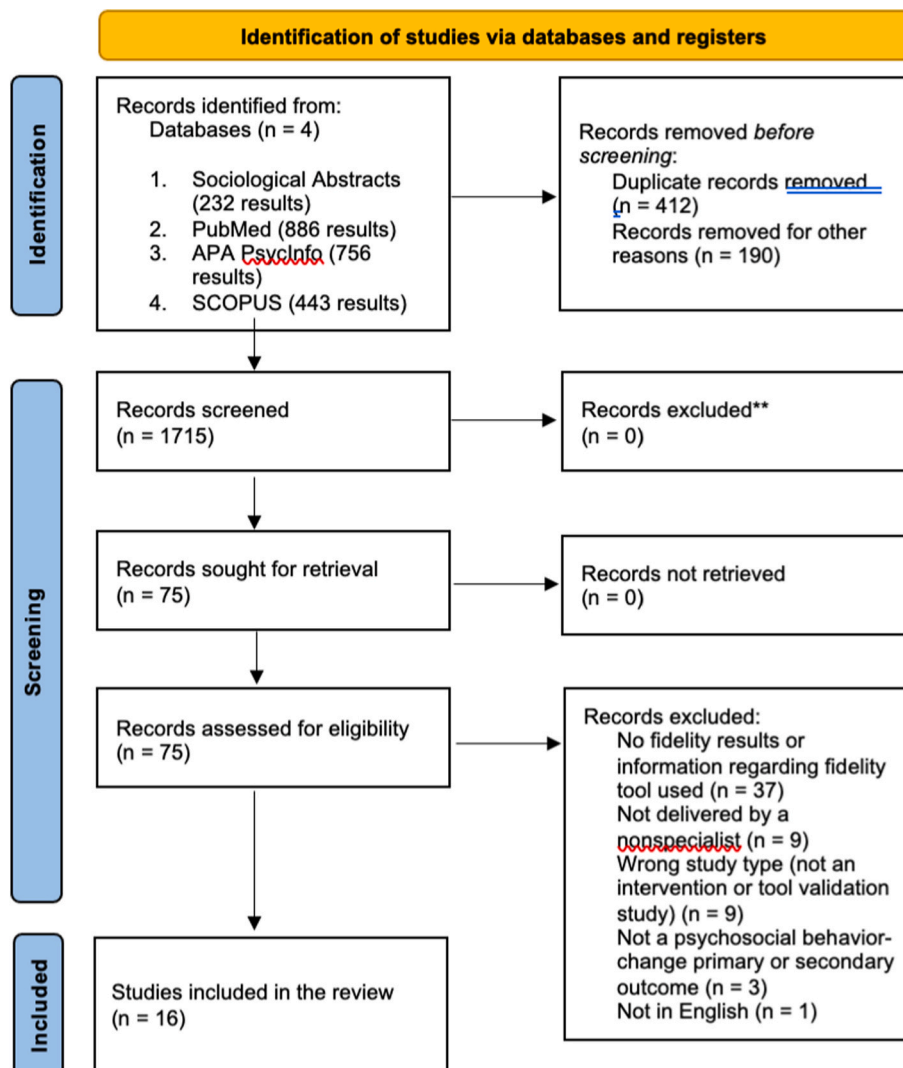
Figure 1. PRISMA flowchart of the screening process and results.

outcome (e.g., mental health), c) studies with information on the tool used to collect fidelity and/or competence data and/or an analysis using such data. We limited our search to studies published after January 1, 2000, as implementation science and an emphasis on effectiveness studies emerged as a field of study in the late 1990s and early 2000s (Bauer & Kirchner, 2020). Finally, we excluded studies that were not published in English, book chapters, and study protocols. When selecting search terms to correspond to each keyword (nonspecialist, evidence-based behavioral intervention, and implementation fidelity), we built search terms around definitions of each concept established in implementation science literature (see Key Concepts, p. 3, for concept definitions).

2.2. Study selection

First, the first and second authors screened all studies according to an eligibility checklist (see Appendix 1 for title abstract screening tool). Two authors reached a 90% inter-rater reliability (IRR) during abstract screening (on 10 abstracts), which increased to 100% after discussion and assessment of an additional 10. Authors split the remaining abstracts to screen separately and compared results at the end. In cases of a discrepancy or a question regarding study inclusion, the study was discussed between authors until a consensus was reached. A third author was available for consultation in the case that consensus could not be reached. In the abstract screening phase, studies were most frequently eliminated due to study outcomes that focused on physical health rather than mental health. Other reasons for exclusion included not being a behavior-change intervention, not mentioning facilitation by nonspecialists, or not referencing fidelity or competence.

After completion of a title and abstract screening, authors moved to a full-text screening (see Appendix 2 for full-text screening tool). Initially, two authors screened full texts simultaneously during two hour-long sessions. In these sessions, authors screened a total of 30 studies. The remaining 45 studies were divided equally between both authors for screening, though authors continued to meet weekly to discuss screening questions and challenges. In addition to the criteria in Appendix 3, full texts that focused exclusively on the development or validation of a fidelity and/or competence monitoring tool were included as long as the tool was developed with the intention of being used in behavior-change interventions, the tools were intended for—or could be used by—nonspecialist, and the text reported how the tool should be measured, assessed, or scored. In the full text screening phase, studies were most frequently excluded due to lack of fidelity and/or competence results or no description of fidelity and/or competence outcomes, even though monitoring of these outcomes was often mentioned as part of intervention delivery. Additional reasons for exclusions are included in Fig. 1. After we selected 17 studies for inclusion, two of the studies pertained to the same intervention even though they fit our inclusion criteria: one was a tool validation study and the other was reporting implementation outcomes. We removed the tool validation study.

2.3. Data extraction

Authors considered the following study characteristics when extracting data from full text screening of manuscripts: 1) tool(s) used to measure or monitor fidelity and/or competence throughout the intervention, 2) quantitative and/or qualitative methods used to analyze fidelity and/or competence data post-intervention, 3) study setting (including country, socio-economic status of the intervention community, rural vs. urban, any other unique characteristics), 4) the roles of supervisors and professionals, 5) type of intervention and intervention intended outcomes, and 6) characteristics of nonspecialist facilitators (see Appendix 4). Authors used Covidence—a systematic review extraction and screening software—for full-text screening and data extraction (Cochrane Community, n.d.).

The first two authors contributed to development of data extraction

categories and screened a total of 16 included studies. Data extraction categories were guided by the research questions with the intention of allowing authors to synthesize information regarding tools and methods that have been used to collect fidelity and/or competence data, and ultimately, understand how the concepts of fidelity and competence have been measured in evidence-based, psychosocial and behavioral interventions. Before authors began independent screening, the data extraction template was piloted and established. Authors first screened the same three studies together, any discrepancies between authors were discussed until consensus was reached. Then, authors divided the remaining studies for independent data extraction.

2.4. Study quality assessment

We assessed the quality of studies and risk of bias using an adapted Mixed Methods Appraisal Tool (MMAT), which has been used widely in systematic reviews assessing the appropriateness of collected data to the stated research question. This was done simultaneously with data extraction, using Covidence (Cochrane Community, n.d.). The tool has been recently revised and includes criteria on five categories of studies, including mixed methods (Hong et al., 2019). We opted to modify the MMAT to better align with our research purposes. More precisely, the adapted MMAT used for this systematic review offers criteria that are capable of evaluating specific subcomponents of methodology the pertain directly to fidelity elements. While the original MMAT focuses on methods (e.g., randomization and sample) and data strictly for intervention outcomes, the adapted MMAT is capable of assessing studies reporting fidelity findings with qualitative, quantitative, or mixed methods data, and provides response options of yes, no, or can't tell for a series of questions. Additionally, we have added a new section to the adapted MMAT that directly addresses fidelity and/or competence measurements, methods, and analysis.

New guidelines from Hong and colleagues (2018) discourage the calculation of total or summative scores for the MMAT. The authors advise to offer more nuanced detail of findings to inform study quality. Each author independently evaluated each article, and in situations of a discrepancy, discussed each article in line with the guidelines until a consensus had been reached. See Appendix 3 for methodological quality criteria included on the MMAT.

2.5. Data synthesis

We used narrative synthesis to report on the findings of our review. To efficiently communicate information, we created two separate tables to synthesize and report the information and data from our included studies. Table 2 summarizes the processes and methods used to monitor fidelity during nonspecialist-delivered interventions. Table 3 summarizes the tools and methods used to analyze fidelity data post-intervention.

3. Results

After performing initial keyword searches in each database, a total of 1715 studies were found after removing duplicates. Each of the 1715 studies were screened according to a title and abstract screening tool (see Appendix 1), and 75 were included in a subsequent full-text screening (see Appendix 2). We screened the 75 full texts and ultimately included 16 studies in our final review.

3.1. Supervising and monitoring

Table 2 offers information on the geographic location, characteristics of interventions, and their objectives. The interventions evaluated in this systematic review were implemented all over the world—with studies deriving from North America, Africa, and Asia. The settings and target populations for each intervention were similarly diverse.

Table 2
Data monitoring.

Study ID	Study	Country	Study Description	Nonspecialist Facilitators	Supervisors and Professionals
1	Asher et al., (2021)	Ethiopia	Community-based rehabilitation for people with schizophrenia in a rural area	Laypersons had at least a tenth- grade education and most had experience with community-based work (equal gender split, mean age 23, age range 20–37)	Two supervisors were assigned to five laypersons each. Supervisors made regular risk assessments of the home visit environment and rated laypersons during routine home visit sessions according to the ENACT tool
2	Atif et al. (2019)	Ethiopia	Bi-monthly group therapy sessions (18 booster sessions) for new mothers from birth of their child up until age three	Peer volunteers were women of child-bearing age paired with participants with similar life circumstances (e.g., socioeconomic status or having experienced perinatal depression)	Peer volunteers were trained and supervised by nonspecialist facilitators from an earlier phase of the intervention, who were previously trained and supervised by a specialist. Supervisors met with facilitators monthly and supervised therapy sessions to assess fidelity according to the Quality and Competence Checklist.
3	Cross et al. (2015)	United States	A school-based intervention to strengthen emotion regulation-skills	Trained paraprofessionals who had held positions in schools such as classroom aides	A trained team of supervisors, including two of the intervention developers, assessed fidelity by coding the videotaped sessions.
4	Diebold et al. (2020)	United States	A perinatal depression home-visiting intervention for new mothers in low-income areas	Female community members were self-selected or selected by a supervisor, and did not have advanced degrees	A single PI provided supervision for facilitators, over the phone in a group setting, during a facilitator's implementation of her first cohort (six group sessions)
5	Garber-Epstein, Zisman-Ilani, Levine, and Roe (2013)	Israel	The Illness Management and Recovery (IMR) Program, an evidence-based, psycho-social intervention for those with serious mental illnesses	Peers (former intervention participants) and paraprofessionals who participated in two days of IMR training	Ongoing supervision throughout intervention delivery
6	Johnson et al., (2021)	United States	A motivational interviewing (MI) intervention for veterans with post-traumatic stress symptoms	Peers (veterans who had also suffered from post-traumatic stress symptoms)	The first author provided supervision of nonspecialists on a weekly individual and/or group supervision (based on staffing, scheduling, and caseloads over the course of the study). In the supervision sessions, feedback was provided based on a review of audio-recorded sessions.
7	Jordans et al., (2021)	Palestine	A group psycho-social intervention for teens and preteens in the Gaza strip who had been exposed to trauma	Nonspecialist facilitators (N = 25; 76% female; mean age = 24.6) were selected at random from a group of trainees recruited to be trained as psychosocial service providers for ongoing programs by War Child Holland	N/A – Tool validation
8	Khan et al. (2019)	Pakistan	A *trans*-diagnostic intervention for women with common mental disorders	Female nonspecialists with at least 16 years of education (graduates) and no formal training in mental health	An apprenticeship model was used in which trained psychiatrists and psychologists built the skills of nonspecialists through on-the-job training. Training included an initial skills training and then four weeks of practice cases with weekly group supervision.
9	Kohrt et al. (2015)	Nepal	An initiative to improve mental health care in primary care settings	Primary care workers being trained in psychological treatment through the intervention	N/A – Tool validation
10	Landry et al. (2019)	United States	A cognitive instruction program for dual-language learners in elementary schools with social and behavioral outcomes	Parents of children similar to intervention participants	Coaches delivered program training to both parents (paraprofessioals) and teachers (professionals) delivering the intervention and supported with development of weekly lesson plans. Coaches also monitored fidelity through audio-recorded sessions.
11	Laurenzi et al. (2020)	South Africa	A home-visiting maternal health care intervention with a depression outcome	Peers (local mothers of participants) who were identified as "positive deviants" – mothers who were able to rise above adversity and raise healthy children	Master Trainers provided training for Mentor Mothers
12	Mastroleo, Mallett, Turrisi, and Ray (2009)	United States	An intervention using motivational interviewing (MI) techniques to reduce college drinking	Peers (undergraduate students) of intervention participants with limited, previous exposure to MI	A Clinical Psychologist and a doctoral student in Counselor Education and Supervision provided weekly individual and group supervision, and assessed fidelity.
13	Munodawafa, Lund, and Schneider (2017)	South Africa	A psychosocial intervention for perinatal depression	Community Health Workers with previous experience doing health promotion visits with mothers and their children under five years	Supervisors were mental health counselors with a Master's in Clinical Social Work and additional supervision was provided by a Senior Clinical Psychologist. Supervisors provided weekly, in-person, group supervision sessions and individual supervision. Supervisors also assessed fidelity by direct observation and through audio recordings.
14		Kenya	A family therapy intervention		

Table 2 (*continued*)

Study ID	Study	Country	Study Description	Nonspecialist Facilitators	Supervisors and Professionals
	Puffer, Friis-Healy, Giusto, Stafford, and Ayuku (2021)			Nonspecialists who self-identified as a religious leader or policy maker.	Supervisors were medical psychology undergraduate students who met with nonspecialists over the phone and in person. These supervisors were also supervised by clinical psychologists on the team
15	Rahman et al. (2019)	Pakistan	A perinatal depression program in conflict-affected area	Lady Health Workers (community health workers employed by the government)	Specialist Lady Health Workers led monthly, in-person group supervision sessions and provided feedback to nonspecialists while troubleshooting challenges with them.
16	Singla et al., (2020)	India	A peer-delivered perinatal depression RCT	Peers of mothers (belonging to the same or neighboring community, with similar socio-demographic backgrounds and good communication skills)	Expert therapists met with nonspecialists in group supervision sessions and rated fidelity via randomly-selected audio recordings of sessions

Interventions were delivered in central group community locations, hospitals or health-based centers, individual residences, refugee sites, schools, and beyond. Additionally, the groups targeted varied from children to adults, and people suffering from mental illness to those without access to adequate mental health services.

Table 2 also describes the processes of monitoring nonspecialist fidelity and competence throughout the intervention and the roles of supervisors and nonspecialists. All studies included specialists as supervisors or assessors of nonspecialist fidelity and competence within the intervention. The roles of supervisors and professionals in the studies were similar and focused on collection of data and monitoring of nonspecialists, in addition to mentorship and support roles. In intervention studies that assessed fidelity and competence of nonspecialists as they delivered the intervention, specialists provided mentorship to nonspecialists by regularly meeting with them one-on-one. Supervisors often reviewed notes or recordings from previous sessions that nonspecialists delivered in order to provide specific feedback for nonspecialists (Asher et al., 2019; Atif et al., 2019; Johnson et al., 2021; Garber-Epstein et al., 2013; Puffer et al., 2021; Singla et al., 2020).

The amount of, and modality of, supervision that nonspecialists received varied between studies. Two studies provided examples of supervisors using didactic role play methods to help nonspecialists prepare for upcoming sessions (Rahman et al., 2019; Puffer et al., 2021). Supervision took place either in-person (Singla et al., 2020), remotely, via phone or video call (Diebold et al., 2020), or both in-person and remotely (Puffer et al., 2021; Rahman et al., 2019). Supervision sessions typically took place on a weekly basis due to the short-term nature of the interventions, though in one intervention delivered over the course of nine months, supervision sessions took place bi-monthly (Garber-Epstein et al., 2013).

Most supervision sessions were held in small groups (Garber-Epstein et al., 2013; Landry et al., 2019; Khan et al., 2019; Rahman et al., 2019; Singla et al., 2020). Reported benefits of small group meetings included collective problem-solving and learning from the experiences of other nonspecialists (Khan et al., 2019). Other supervision modalities incorporated a blended method of in-person and group supervision sessions due to work constraints or in order to supplement the benefits of group learning with tailored, individualized feedback to nonspecialists (Puffer et al., 2021; Garber-Epstein et al., 2013; Johnson et al., 2021; Munodawafa et al., 2017). The ratio of specialists to nonspecialists ranged from 1:5 to 1:31 (Asher et al., 2019; Atif et al., 2019; Garber-Epstein et al., 2013; Johnson et al., 2021; Landry et al., 2019; Munodawafa et al., 2017; Puffer et al., 2021; Singla et al., 2020). In these studies, supervisors included the principal investigator (Diebold et al., 2020), peers of intervention participants (Johnson et al., 2021; Rahman et al., 2019; Singla et al., 2020), undergraduate students in medical psychology who were in turn mentored by a clinical psychologist (Puffer et al., 2021), and clinical social workers with a Master's degree (Munodawafa et al., 2017).

The extraction of study characteristics revealed that audio- and video-recorded sessions were valuable for rating the fidelity and competence of nonspecialists, and allowed for data to be revisited post-intervention. In intervention studies, supervisors filled out fidelity checklists post-hoc by reading verbatim transcripts from recordings of sessions (Puffer et al., 2021; Singla et al., 2020; Munodawafa et al., 2017). In a school-based intervention designed to strengthen emotion-regulation skills, videotaped sessions were coded by a trained team of supervisors including two of the intervention developers (Cross et al., 2015). In this intervention, periodic reliability checks were conducted by the research team, in which fidelity and competence checklists filled out by supervisors were simultaneously filled out by a highly-trained member of the research team after watching the video-taped session (Cross et al., 2015). In other interventions, fidelity and/or competence was monitored as trained specialists, who had often previously delivered the intervention, filled out fidelity checklists of nonspecialists during combined supervisory and monitoring visits (Garber-Epstein et al., 2013; Johnson et al., 2021; Landry et al., 2019). In tool validation studies, specialists assisted with the training of nonspecialists and rated their competence on the respective tools (Jordans et al., 2021; Kohrt et al., 2015; Mastroleo et al., 2009; Singla et al., 2020).

3.2. Tools and data analysis

Six studies reported use of a tool that measured competence only, and not fidelity (Asher et al., 2021; Atif et al., 2021; Diebold et al., 2020; Laurenzi et al., 2020 ; Mastroleo et al., 2009; Rahman et al., 2019). Conversely, three studies reported use of a tool that measured fidelity without measuring competence (Garber-Epstein et al., 2013; Landry et al., 2019; Munodawafa et al., 2017). The remaining five studies used tools that measured both competence and fidelity, or else used two separate tools to measure both constructs (Puffer et al., 2021; Singla et al., 2020; Cross et al., 2015; Johnson et al., 2021; Khan et al., 2019).

Most studies reported the use of descriptive statistics to measure average changes in fidelity and/or competence scores at different points in time. Six studies used quantitative, descriptive methods to analyze data, through calculating mean or summative scores from the tool, often comparing scores with a certain threshold that indicated satisfactory fidelity or competence (Asher et al., 2021; Atif et al., 2019; Diebold et al., 2020; Landry et al., 2019; Laurenzi et al., 2020; Rahman et al., 2019). Overall, the sample sizes in these studies were generally small for analyses (ranging from N = 10 to N = 45), which justified the use of descriptive statistics rather than more complex modeling. Some studies offered nuanced changes to fidelity scores and satisfactory status over time (e.g., prior to intervention through to after the intervention). As an example, one study (Asher et al., 2021) showed that empathy competencies improved throughout the intervention the most, while problem-solving and advice-giving competencies saw the least improvements.

Three studies used mixed methods or qualitative approaches to

Table 3

Data analysis.

Study ID	Study	Construct Measured	Tools Used	Methods Used to Analyze Fidelity Data Post-Intervention	Details and Findings of Fidelity Outcomes
Intervention Studies					
1	Asher et al., (2021)	Competence	Ethiopian adaptation of the ENhancing Assessment of Common Therapeutic factors (ENACT) structured observational rating scale	For each time point, mean items scores were generated for each CBR worker, then double-rated competence assessments were averaged. Summary means were generated for each time point and for role play assessments.	Mean scores showed improvement in CBR worker competence throughout the training and the intervention. Empathy scores showed earliest improvements, and problem-solving and advice-giving saw the least improvements. More supervision by specialists was needed.
2	Atif et al. (2019)	Competence	Quality and Competence Checklist, an observational tool used by trainers to rate a group session on 6 areas of competencies	Each area of the fidelity tool was scored on a Likert scale (0–2), ranging from "not demonstrated" to "partially demonstrated" and "demonstrated well", with an option of not applicable, and then converted to a percentage. A percentage of 70% indicated competence.	All 31 of the 45 peer facilitators who were retained over five years, all of them achieved satisfactory competence. Six of the 14 peers who dropped out did so because they could not achieve satisfactory competence.
3	Cross et al. (2015)	Fidelity and Competence	Intervention-specific tool measuring both adherence and competence	An exploratory factor analysis (EFA) was used to examine the factor structure an intra-class correlation was used to measure inter-rater reliability of the tool. Descriptive statistics were used to summarize adherence and fidelity, and multilevel analyses validated that implementer fidelity measures were clustered around the implementer rather than attributable to other factors.	Variance in fidelity scores was explained by the implementer, and intra-class correlations were satisfactory. The EFA revealed two domains of the tool: adherence and competence. Summative adherence and competence scores varied widely and predicted children's enhanced response to the intervention, but not externalizing behavior.
4	Diebold et al. (2020)	Competence	Revised Cognitive Therapy Rating Scale	Descriptive statistics and linear mixed models were used to examine average competence scores and adherence, including fixed effects for study arm and session number. Models also examined site, facilitator, and client-specific effects.	There were no differences between paraprofessionals and professionals for overall adherence or competence. Surprisingly, facilitators with a Master's degree or higher had lower average adherence, and facilitators who were trained via audio recording rather than 1-on-1 had lower average adherence.
5	Garber-Epstein et al. (2013)	Fidelity	The Illness Management and Recovery Fidelity Scale	Analysis of variance (ANOVA) was used to determine mean differences between clinicians delivering the intervention and two groups of nonspecialists (trained peers and other nonspecialists).	Each group of facilitators achieved satisfactory fidelity, with other nonspecialists (not peers) receiving the greatest improvement in mean fidelity scores between timepoint 1 and timepoint 2.
6	Johnson et al., (2021)	Fidelity and Competence	Intervention-specific tool measuring three domains: coaching skills, intervention stages and phases, and peer role	Qualitative interviews were thematically coded. Descriptive statistics and count of prevalence were used to analyze quantitative data from the fidelity tool.	Nonspecialists delivered the intervention with fidelity in more than 90% of sessions.
8	Khan et al. (2019)	Fidelity and Competence	Intervention-specific fidelity tool that measured both competence and fidelity, conceptualized as "counseling skills" and intervention strategies	N/A	Competence scores were low at first for nonspecialists, but with subsequent training and supervision, nonspecialists improved.
10	Landry et al. (2019)	Fidelity	Teacher Behavior Rating Scale-Bilingual Version (TBRS-B)	Descriptive statistics (frequency and percentage scores) from the TBRS-B, which was rated on a Likert scale from 1 to 6	TBRS-B scores increased more for professionally-trained intervention teachers and nonspecialist teachers compared to the control group.
11	Laurenzi et al. (2020)	Competence	Home Visitor Communication Skills Inventory (HCSI) with three domains measuring competence: active delivery, active connecting, and active listening	Descriptive statistics (proportions, frequencies) and correlations between average visit duration, and active delivery and active connecting	Nonspecialists had higher scores in active listening and active delivery than in active connecting.
12	Mastroleo et al. (2009)	Competence	Peer Proficiency Assessment (PEPA)	Correlations computed between nonspecialist and specialist coder scores to examine inter-rater reliability, the PEPA questions and MI adherent scores to examine construct validity, between PEPA scores and effectiveness outcomes (drinking behaviors) to examine predictive validity.	PEPA scores indicated MI adherence (r – 0.872). Assessments also revealed high inter-rater reliability between student and master coders and good correlations between previously established fidelity tools.
13	Munodawafa et al. (2017)	Fidelity	Intervention-specific fidelity tool	Descriptive statistics (mean fidelity scores per session supplemented with key informant interviews	On average, nonspecialists achieved moderate to good intervention facility. Qualitative interviews revealed that the manual and ongoing and training and supervision served as facilitators to achieving intervention fidelity
14	Puffer et al. (2021)	Fidelity and Competence	Intervention-specific fidelity tool guided by the ENACT scale	Descriptive statistics and visual plotting of fidelity and competence ratings to	Nonspecialists achieved adequate fidelity scores. The highest competence score was

Table 3 (*continued*)

Study ID	Study	Construct Measured	Tools Used	Methods Used to Analyze Fidelity Data Post-Intervention	Details and Findings of Fidelity Outcomes
				explore patterns of change across steps of the intervention and variability on specific competencies. Inductive coding of focus groups data and card sorting methods.	structured problem exploration and the lowest was cognitive behavioral skills for children, which are least frequently used.
15	Rahman et al. (2019)	Competence	ENhancing Assessment of Common Therapeutic factors (ENACT) structured observational rating scale	Descriptive statistics (mean scores) were generated, and mean differences in scores for nonspecialists trained virtually and in-person were generated	There were no significant differences in scores between groups of nonspecialists
16	Singla et al., (2020)	Fidelity and Competence	Therapist Quality Scale, measuring both fidelity and competence (treatment-specific and general skills)	Assessment of inter-rater reliability (intraclass correlation coefficients), internal consistency, and predictive validity of patient outcomes (depression)	There were moderate to excellent scores of inter-rater reliability among specialists (ICC = 0.779) and nonspecialists (ICC = 0.714); there was high internal consistency (α = 0.814 for specialist coders and α = 0.843) for nonspecialist coders, and TQS ratings were not significantly related to clinical outcomes (r = 0.375, p < 0.01).
Tool Development and Validation Studies					
7	Jordans et al., (2021)	Competence	WeACT instrument, which was modeled after ENACT	Assessment of inter-rater reliability (intraclass correlation coefficients) and internal consistency.	At timepoint 1 (N = 8 raters), ICC = 0.47 (95% C.I. 0.26–0.72), α = 0.91; At timepoint 2 (N = 6 raters), ICC = 0.68 (95% C.I. 0.48–0.86), α = 0.94
9	Kohrt et al. (2015)	Competence	ENhancing Assessment of Common Therapeutic factors (ENACT) structured observational rating scale	Assessment of inter-rater reliability (intraclass correlation coefficients)	ICC = 0.88 for experts (95% C.I. 0.81–0.93); ICC = 0.67 (95% CI 0.60–0.73) for nonspecialists.

examine fidelity and competence of nonspecialist facilitators (Johnson et al., 2021; Munodawafa et al., 2017; Puffer et al., 2021). These studies used convergent mixed methods designs and complemented descriptive quantitative data from fidelity and/or competence checklists with qualitative data collected from participants, supervisors, or non-specialists themselves (Johnson et al., 2021; Munodawafa et al., 2017; Puffer et al., 2021). Johnson and colleagues supplemented data from the fidelity checklist, which was analyzed by descriptive statistics, with data from qualitative semi-structured interviews from key informant participants. This design also allowed for the triangulation of data – quantitative data was collected from specialists observing the fidelity of nonspecialist facilitators, and qualitative data was collected from participants receiving the intervention (a peer coaching intervention to address post-traumatic stress symptoms in veterans) from nonspecialist facilitators (peers of veterans). Qualitative data captured perceptions of the coaching (including how well it fit their needs), helpfulness, suggestions for improvement, the relationship with the peer coach, and intervention intensity (Johnson et al., 2021).

Only three studies used quantitative methods beyond descriptive statistics; one of these studies was limited by a small sample size due to small numbers of nonspecialist providers. Diebold et al.'s (2020) study used linear mixed models to examine average competence scores and adherence, yet notes the small sample size as a significant limitation. Garber-Epstein and colleagues assessed differences in mean fidelity scores between clinicians delivering a psychosocial intervention and two groups of nonspecialists across two time points (trained peers of participants and other nonspecialists who were trained to deliver the intervention), using analysis of variance (ANOVA) (2013). Compared to other fidelity analyses, the Garber-Epstein study had a larger sample size, (N = 210) facilitators across the three groups. Cross and colleagues used multilevel modeling to examine the degree to which variance in nonspecialist fidelity measures was attributed to the individual nonspecialist and to the participant receiving the intervention. This method was used to validate that fidelity measures accurately reflected nonspecialist performance, not other variables which were accounted for by fixed effects or by the participant (Cross et al., 2015). Table 3 describes the methods used to analyze fidelity data that was collected throughout an intervention.

3.3. Methodological quality of included studies

Studies included in the review demonstrated adequate methodological quality according to the adapted MMAT. Out of the 16 studies assessed, 13 were rated as high quality, suggesting that the methods used were adequate to answer the research question and that the studies were rigorous and detailed in the descriptions of their methods and findings. One study (Diebold et al., 2020) used linear mixed models with a low-powered study and fidelity and/or competence outcomes were not operationalized. Three studies (Johnson et al., 2021; Khan et al., 2019; Landry et al., 2019) did not provide sufficient detail when reporting on the fidelity and/or competence findings, and both Khan et al. and Landry et al. could have operationalized outcomes better and provided more detail regarding the collection of fidelity and/or competence ratings.

4. Discussion

The purpose of our systematic review was to identify the tools and processes being used to monitor fidelity and/or competence of evidence-based behavioral interventions, which is critical to task-shifting, quality delivery in pursuit of pre-identified effects, and sustaining evidence-based, behavioral interventions. Implemented interventions that diverge significantly from their intended delivery (i.e., are delivered with low fidelity), may no longer qualify as evidence-based. Additionally, if quality of delivery is not maintained, task-shifting to nonspecialists may no longer serve as an effective strategy that bridges the treatment gap in low-resource settings. As we consider the results from our data extraction within existing evidence from the broader field of implementation research, our systematic review provides the following conclusions:

1. Inconsistency in methods, reporting standards, operationalization, and language is a challenge; common terminology and expectations are important for comparison.
2. Supervision, leadership, and coaching play a strong role in fidelity or competence maintenance, improvement, management, and evaluation.
3. Methods, measures, and analysis plans for fidelity or competence assessment will usually be conditional on the program being

evaluated, but there is room for improvement. Prototypes may not be desirable for fidelity tools, as the tools measure adherence to the intervention manual. However, competence tools measure common elements that are shared across behavior-change interventions, and establishing tool prototypes can establish best practices and strengthen the validity and reliability of measures.

4.1. Fidelity vs. competence

Results indicated that some studies measured fidelity of non-specialists without attention to competence, and some measured competence without attention to fidelity. Only five of the 16 included studies used a tool that measured both fidelity and competence, though the literature indicates that this is critical for ensuring the quality of delivery (Cross et al., 2015; Johnson et al., 2021; Khan et al., 2019; Puffer et al., 2021; Singla et al., 2020).

Findings from two meta-analytic reviews offer insight into the distinction and selection of competence instruments and fidelity instruments (Webb, Derubeis, & Barber, 2010; Collyer, Eisler, & Woolgar, 2019). These reviews, which stratified their analyses to distinguish fidelity and competence as separate outcomes, concluded that competence had weak associations with intervention effectiveness outcomes. However, both reviews acknowledged weaknesses in the measurement of competence (Webb et al., 2010; Collyer et al., 2019). Perez and colleagues (2019) also discusses the distinction between competence and fidelity, but refer to these terms as core functions (fidelity) and forms (competencies). Their findings contradict results from the meta-analytic studies. While fidelity is assessed at the function level, Perez and colleagues argue that the success of the intervention is contingent upon the facilitator (and the intervention implementers at large) tailoring each function to the specific needs and the specific context of the intervention and its participants (2019). The importance of context is a critical component of competence (Waltz et al., 1993). While fidelity, as defined above, focuses on adherence to the manual and the intervention protocol, competence more broadly refers to the "level of skill shown by the therapist in delivering the treatment" and "the extent to which the therapists conducting the interventions took the relevant aspects of the therapeutic context into account and responded to these contextual variables appropriately" (Waltz et al., 1993, p. 620).

Assessments of fidelity and competence each provide different information to the research team, and it is crucial to pay attention to both fidelity to the manual and awareness of contextual factors and common therapeutic attributes. Accurate measurement of competence could allow interventions to be adapted in real-time to best fit the needs of the clients, considering all ecological factors at play, and thus leading to results that are more widely generalizable within diverse global health contexts. In a call to action regarding emerging opportunities in global health, Theobold et al. complement arguments (Perez et al., 2019; Waltz et al., 1993) that intervention success often requires tailoring each function to the specific needs and the specific context of the intervention and its participants (Theobald et al., 2018). Without reference to "competence" or "form," they illustrate how, at times, a tension exists between the maintenance of intervention fidelity and the need to be able to adapt the intervention throughout the course of implementation (Theobald et al., 2018). Ultimately, adaptations can improve the effectiveness of EBIs when facilitators are trained to recognize different contexts and needs (Theobald et al., 2018). Murray et al. refer to the "flexibility within fidelity" that provides space for creativity and adaptation to account for context and ensure better intervention fit to the population (2011). This approach supports the idea that the concept of fidelity cannot supersede competence, nor can competence supersede fidelity. Nonspecialists must be equipped with sufficient knowledge of the intervention and its manual to understand when, and how, to move beyond the manual and to deliver content in a contextually-appropriate manner that remains faithful to the overall purpose of the intervention

while still meeting the needs of clients and participants.

4.2. Value of quality supervision and leadership

The presence and responsibilities of supervisors highlights the importance of professionals and emphasizes that capacity building and training of nonspecialists happens throughout the course of the intervention. Supervision is critical to non-specialist delivered interventions as it provides a space for structured and reliable feedback, and develops and maintains competence (Kohrt et al., 2015; Singla et al., 2020). While pre-intervention training is certainly critical for equipping non-specialists, nonspecialists continue to grow and learn as the intervention is delivered. Problem-solving was listed on multiple tools, and thus, regular meetings with supervisors is helpful for troubleshooting issues that arise during intervention delivery (Asher et al., 2021; Atif et al., 2019; Jordans et al., 2021; Kohrt et al., 2015; Laurenzi et al., 2020; Singla et al., 2020). The inclusion of specialists in interventions is also beneficial for ensuring the quality of data collection, as specialists are best equipped to identify if an intervention is being delivered according to its original intention and with high quality (Singla et al., 2018). Specialists were often assigned to complete fidelity and/or competence checklists throughout the course of the intervention by either directly attending sessions or by observing video-recordings of sessions. In this manner, specialists were able to amplify their influence and expertise throughout the intervention by monitoring nonspecialists and serving as supervisors and mentors of these nonspecialists. Supervising professionals are instrumental in both training nonspecialists to deliver an intervention with quality (fidelity and competence), and assessing quality throughout the intervention.

4.3. Method and data analysis in fidelity monitoring

A key observation we made throughout the process of compiling and analyzing our findings was the range of methodologies used to measure fidelity and/or competence. A majority of the studies included in this systematic review used a form of descriptive statistic reporting to evaluate fidelity (Asher et al., 2019; Atif et al., 2021; Cross et al., 2015; Diebold et al., 2020; Garber-Epstein et al., 2013; Landry et al., 2019; Laurenzi et al., 2020; Rahman et al., 2019), though a handful of studies used qualitative or mixed-methodologies (Johnson et al., 2021; Muno-dawafa et al., 2017; Puffer et al., 2021), and one study used a linear model (Diebold et al., 2020). We attribute this variety of research designs and tools, and lack of inferential statistics, to the novelty of fidelity instruments and the lack of large sample sizes for lay worker workforces who are delivering interventions. It is possible that the diversity of tools and methods used is not only warranted, but optimal for the exercise of fidelity and/or competence monitoring. It is not unreasonable to imagine that tracking descriptive statistics or count observations or simple supervisor observation is all that is necessary to evaluate fidelity and/or competence, and is more straightforward for community-based programs to conduct. Perhaps more advanced statistical models and protocols are not required to merit trust in quality of intervention delivery. Furthermore, analyses using descriptive statistics are more accessible for local partners or agencies implementing evidence-based interventions without advanced methods training, and therefore, simpler methods can help bridge the gap between research and practice. We must consider the core purpose of fidelity and competence monitoring—delivering evidence-based interventions as intended. Our research questions regarding implementation and quality delivery should meet a minimally sufficient threshold to offer enough confidence that programs are being delivered as intended.

In the future, we suggest that more precise terminology be used by both researchers and practitioners across the implementation science and fidelity monitoring literature. This is vital to make key distinctions between concepts like fidelity, competence, adherence, and other terms frequently used interchangeably. Further, we recommend that future

research focuses on the impact of both competence and fidelity on effectiveness outcomes for programs being delivered. Additionally, model development and framework specification may be necessary to determine a valid definition of fidelity and its subcomponents. This will help clarify whether competence scoring tools are sufficient to capture fidelity, or whether more comprehensive tools are required. We also suggest that practitioners and researchers doing implementation studies use expert consultation to generate guidelines on how supervisors can adhere to best practice when seeking to train, monitor, and provide feedback in regards to fidelity. As seen in this review, supervisors are a valuable asset to the structured delivery of high-quality interventions. Finally, we suggest deeper investigation into methods and data analysis necessary for fidelity assessment to establish a minimal threshold. This review revealed that there is still a large degree of variance in what counts as sufficient fidelity analysis. Identifying commonalities in analysis plan, establishing minimal standards, and creating a repository of fidelity monitoring batteries or case examples could help develop methodological guidelines for best practice. Platforms such as EQUIP or EMPOWER, and guidance from mhGAP, can be leveraged for both training and supporting nonspecialists throughout the course of intervention delivery. Fidelity tools in evidence-based, behavioral interventions are vital to the systematic progression and implementation of evidence-based behavioral interventions.

4.4. Limitations

This systematic review presented several challenges. First, a lack of common language and terminology. This produced concerns in regard to comparing different methods and tools, as well as identifying fidelity and/or competence focused peer-reviewed papers. Certain papers may have remained undetected if they did not directly mention fidelity or competence explicitly, even if they had measured these concepts. We sought to overcome this by using an inclusive net of keywords that we incorporated into our search terms protocol, to offer a comprehensive snapshot of what is currently being done.

A second issue that arose were the diversity of methodologies in data analysis, tools, and protocols to assess fidelity. Given that implementation science is still in its infancy, this to an extent is unavoidable and protocols should be using contrasting measurement systems to match their designs. Nevertheless, the lack of common foundation or best practice guidelines makes comparisons across tools and study designs difficult. We attempted to overcome this by adapting a commonly used systematic review quality assessment tool. The original tool was substantially oriented toward assessing the quality of study designs. This was not suitable for our current needs given the sheer variety of study types and configurations. But with our adaptations the tool was sufficient for assessing the fidelity and/or competence instruments and protocols in each included study. We are hopeful the adapted MMAT will be useful in future evaluations of fidelity and/or competence.

Finally, our review may be limited by publication bias and unintentional remission of valuable data regarding fidelity and competence monitoring and tools. While intervention studies often include fidelity and competence components, and measure both of these concepts using validated tools, this information is often used for internal purposes and not reported on in the peer-reviewed literature. We hope that our review encourages researchers and practitioners to publish more implementation-related literature related to fidelity and competence tools and data analysis.

5. Conclusion

Using similar tools to measure fidelity and/or competence in evidence-based, behavioral interventions is beneficial as findings can then be compared across interventions. In addition, intervention-specific fidelity items (such as adherence to a tailored curriculum) can easily be added onto existing tools with a base of interpersonal skills (competence). These skills are relevant regardless of the population being served by the intervention. Modalities involved integration of mental health care into primary health care settings (Asher et al., 2021; Kohrt et al., 2015), group psychotherapy sessions (Jordans et al., 2021; Atif et al., 2019), and a home-visiting intervention (Diebold et al., 2019).

CRediT authorship contribution statement

Laura Bond: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Visualization, Writing – original draft. **Erik Simmons:** Methodology, Validation, Formal analysis, Investigation, Visualization, Writing – original draft. **Erika L. Sabbath:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no competing interests.

Data availability

No data was used for the research described in the article.

Appendix 1. Title and Abstract Screening Tool

Is the intervention a behavior-change intervention? *	Abstract or title must include this criterion
Does the study mention facilitation by nonspecialists?	Abstract or title must include at least one of these criteria
Does the study mention fidelity and/or competence monitoring, a fidelity and/or competence tool, a fidelity and/or competence outcome, or an assessment of nonspecialist competence or facilitation quality?	

*DOES INCLUDE behavior-change interventions with mental health and psycho-social outcomes (can include clinically diagnosed mental health conditions, can include a psychosocial outcome for populations experiencing a physical ailment), and tool validation studies.

DOES NOT INCLUDE physical health outcomes of any kind (such as medicine adherence, cancer screening, breastfeeding practices, nutrition practices, new HIV/AIDs case reductions, weight loss, sexually transmitted diseases, autism, cardiac arrest, pediatric care, etc.), systematic reviews, results from pre-intervention trainings.

Appendix 2. Full Text Screening Tool

Is the intervention a behavior-change intervention?* Was the intervention delivered by nonspecialists? Does the study mention fidelity and/or competence monitoring, a fidelity and/or competence tool, a fidelity and/or competence outcome, or an assessment of nonspecialist competence or facilitation quality?	Manuscript must include these criteria

*DOES INCLUDE behavior-change interventions with mental health and psycho-social outcomes (can include clinically diagnosed mental health conditions, can include a psychosocial outcome for populations experiencing a physical ailment), and tool validation studies.

DOES NOT INCLUDE physical health outcomes of any kind (such as medicine adherence, cancer screening, breastfeeding practices, nutrition practices, new HIV/AIDs case reductions, weight loss, sexually transmitted diseases, autism, cardiac arrest, pediatric care, etc.), systematic reviews, results from pre-intervention trainings.

Appendix 3. Adapted Mixed Methods Appraisal tool

Category of study designs	Methodological quality criteria	Responses			
		Yes	No	Can't tell	Comments
Screening questions (for all types)	S1. Are there clear research questions? S2. Do the collected data allow to address the research questions? *Further appraisal may not be feasible or appropriate when the answer is 'No' or 'Can't tell' to one or both screening questions.*				
1. Qualitative	1.1. Is the qualitative approach (including data collection methods) appropriate to answer the research question? 1.2. Are the findings adequately derived from the data? 1.3. Is there coherence between qualitative data sources, collection, analysis and interpretation?	___	___	___	___
2. Quantitative	2.1. Are measurements appropriate regarding both the outcome and intervention (or exposure)? 2.2. Is adequate detail provided about the data collection process? 2.3. Is the statistical analysis and interpretation appropriate to answer the research question?	___	___	___	___
3. Mixed methods	3.1. Is there an adequate rationale for using a mixed methods design to address the research question? 3.2. Are the different components of the study effectively integrated to answer the research question? 3.3. Are divergences and inconsistencies between quantitative and qualitative results adequately addressed?	___	___	___	___
4. Fidelity (for all studies)	4.1. Is reference made to a validated fidelity and/or competence tool? 4.2. Is there sufficient detail reported on the fidelity and/or competence findings? 4.3. Is information provided on how fidelity and/or competence ratings are collected? 4.4. Are fidelity-related and/or competence-related outcomes operationalized?				

Appendix 4. Data Extraction Form

Study ID
Title   Reviewer Name   Covidence ID   Lead author   Country in which the study was conducted   Study design   Mode of delivery   Treatment setting (physical location)   Total number of intervention participants   Population characteristics of intervention participants   Total number of nonspecialist facilitators   Population characteristics of nonspecialist facilitators   Roles of supervisors and professionals   Effectiveness outcomes   Quantitative and/or qualitative methods used to analyze effectiveness data in the intervention   Implementation outcomes   Tool(s) used to measure or monitor fidelity and/or competence throughout the intervention   Method(s) used to measure or monitor fidelity and/or competence throughout the intervention   Quantitative and/or qualitative methods used to analyze fidelity and/or competence data post-intervention   Details and findings of fidelity and/or competence outcomes   Notes

Appendix 5. Study Identification

1	"Like a doctor, like a brother": Achieving competence amongst lay health workers delivering community-based rehabilitation for people with schizophrenia in Ethiopia	Asher et al., (2021)
2	Delivering maternal mental health through peer volunteers: a 5-year report	Atif et al. (2019)

(*continued*)

3	Observational measures of implementer fidelity for a school-based preventive intervention: development, reliability, validity	Cross et al. (2015)
4	Comparing Fidelity Outcomes of Paraprofessional and Professional Delivery of a Perinatal Depression Preventive Intervention	Diebold et al. (2019)
5	Comparative impact of professional mental health background on ratings of consumer outcome and fidelity in an Illness Management and Recovery Program	Garber-Epstein et al. (2013)
6	Engagement, experience, and satisfaction with peer-delivered whole health coaching for veterans with PTSD: A mixed methods process evaluation	Johnson et al., (2021)
7	Assessment of service provider competence for child and adolescent psychological treatments and psychological services in global mental health: evaluation of feasibility and reliability of the WeACT tool in Gaza, Palestine	Jordans et al., (2021)
8	Evaluating feasibility and acceptability of a group WHO *trans*-diagnostic intervention for women with common mental disorders in rural Pakistan: a cluster-randomized controlled feasibility trial	Khan et al. (2019)
9	Therapist competence in global mental health: Development of the Enhancing Assessment of Common Therapeutic factors (ENACT) rating scale	Kohrt et al., (2015)
10	The effect of the Preparing Pequeños small-group cognitive instruction program on academic and concurrent social and behavioral outcomes in young Spanish-speaking dual language learners	Landry et al. (2019)
11	The home visit communication skills inventory: Piloting a tool to measure community health worker fidelity to training in rural South Africa	Laurenzi et al. (2020)
12	Psychometric properties of the Peer Proficiency Assessment (PEPA): a tool for evaluation of the undergraduate peer counselor's motivational interviewing fidelity	Mastroleo et al. (2009)
13	A process evaluation exploring the lay counselor experience of delivering a task shared psycho-social intervention for perinatal depression in Khaleyitsa, South Africa	Munodawafa et al. (2017)
14	Development and Implementation of a Family Therapy Intervention in Kenya: a Community-Embedded Lay Provider Model	Puffer et al. (2021)
15	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	Rahman et al. (2019)
16	Peer supervision for assuring the quality of nonspecialist provider delivered psychological intervention: Lessons from a trial for perinatal depression in Goa, India	Singla et al., (2020)

References

Asher, L., Birhane, R., Teferra, S., Milkias, B., Worku, B., Habtamu, A., et al. (2021). Like a doctor, like a brother": Achieving competence amongst lay health workers delivering community-based rehabilitation for people with schizophrenia in Ethiopia. *PLoS One, 16*(2), Article e0246158.

Atif, N., Bibi, A., Nisar, A., Zulfiqar, S., Ahmed, I., LeMasters, K., et al. (2019). Delivering maternal mental health through peer volunteers: A 5-year report. *International Journal of Mental Health Systems, 13*, 62.

Barber, J. P., Sharpless, B. A., Sharpless, Klostermann, S., & McCarthy, K. (2007). Assessing intervention competence and its relation to therapy outcome: A selected review derived from the outcome literature. *Professional Psychology: Research and Practice, 38*(5), 493–500. https://doi.org/10.1037/0735-7028.38.5.493

Barnart, D., Farrar, J., Murray, S., Brennan, R., Antonaccio, C., Sezibera, V., et al. (2020). Lay-worker delivered home visiting promotes early childhood development and reduces violence in Rwanda: A randomized pilot. *Journal of Child and Family Studies, 29*, 1804–1817.

Bauer, M. S., & Kirchner, J. (2020). Implementation science: What is it and why should I care? *Psychiatry Research, 283*, Article 112376. https://doi.org/10.1016/j.psychres.2019.04.025

Betancourt, T. S., Hansen, N., Farrar, J., Borg, R. C., Callands, T., Desrosiers, A., et al. (2021). Youth functioning and organizational success for west african regional development (youth FORWARD): Study protocol. *Psychiatric Services, 72*(5), 563–570. https://doi.org/10.1176/appi.ps.202000009

Bolton, P., Bass, J., Neugebauer, R., Verdeli, H., Clougherty, K. F., Wickramaratne, P., et al. (2003). Group interpersonal psychotherapy for depression in rural Uganda: A randomized controlled trial. *JAMA, 289*(23), 3117–3124. https://doi.org/10.1001/jama.289.23.3117

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science, 2*(1), 1–9.

n.d. Cochrane Community. Covidence softward. from https://community.cochrane.org/help/tools-and-software/covidence. (Accessed 17 August 2022)

Collyer, H., Eisler, I., & Woolgar, M. (2019). Systematic literature review and meta-analysis of the relationship between adherence, competence, and outcome in psychotherapy for children and adolescents. *European Child & Adolescent Psychiatry, 29*, 417–431.

Crea, T., Usher, C., & Wildfire, J. (2009). Implementation fidelity of team decision-making. *Children and Youth Services Review, 31*(1), 119–124.

Cross, W., West, J., Wyman, P. A., Schmeelk-Cone, K., Xia, Y., Tu, X., … Forgatch, M. (2015). Observational measures of implementer fidelity for a school- based preventive intervention: Development, reliability, and validity. *Prevention Science: Official Journal of the Society for Prevention Research, 16*(1), 122–132.

Cuijpers, P., Reijnders, M., & Huibers, M. (2019). The role of common factors in psychotherapy outcomes. *Annual Review of Clinical Psychology, 15*, 207–231.

Desrosiers, A., Schafer, C., Esliker, R., Jambai, M., & Betancourt, T. (2021). mHealth-supported delivery of an evidence-based family home-visiting intervention in Sierra Leone: Protocol for a pilot randomized controlled trial. *JMIR Res. Protocol., 10*(2), Article e25443. https://doi.org/10.2196/25443

Diebold, A., Ciolino, J. D., Johnson, J. K., Yeh, C., Gollan, J. K., & Tandon, S. D. (2020). Comparing fidelity outcomes of paraprofessional and professional delivery of a perinatal depression preventive intervention. *Administration and policy in mental health, 47*(4), 597–605.

Elmore, R. F. (1980). Backward mapping: Implementation research and policy decisions. *Political Science Quarterly, 94*(4), 601–616.

Fairburn, C. G., & Cooper, Z. (2011). Therapist competence, therapy quality, and therapist training. *Behaviour research and therapy, 49*(6–7), 373–378.

Garber-Epstein, P., Zisman-Ilani, Y., Levine, S., & Roe, D. (2013). Comparative impact of professional mental health background on ratings of consumer outcome and fidelity in an Illness Management and Recovery program. *Psychiatric Rehabilitation Journal, 36*(4), 236–242.

Ginsburg, L. R., Hoben, M., Easterbrook, A., Anderson, R. A., Estabrooks, C. A., & Norton, P. G. (2021). Fidelity is not easy! Challenges and guidelines for assessing fidelity in complex interventions. *Trials, 22*(1), 372. https://doi.org/10.1186/s13063-021-05322-5

Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., et al. (2019). Improving the content validity of the mixed methods appraisal tool: A modified e-delphi study. *Journal of Clinical Epidemiology, 111*, 49–59. https://doi.org/10.1016/j.jclinepi.2019.03.008. e1.

Johnson, E. M., Possemato, K., Khan, S., Chinman, M., & Maisto, S. A. (2021). *Engagement, experience, and satisfaction with peer-delivered whole health coaching for veterans with PTSD: A mixed methods process evaluation*. Psychological services.

Jordans, M., Coetzee, A., Steen, H. F., Koppenol-Gonzalez, G. V., Galayini, H., Diab, S. Y., et al. (2021). Assessment of service provider competence for child and adolescent psychological treatments and psychosocial services in global mental health: Evaluation of feasibility and reliability of the WeACT tool in gaza, Palestine. *Global mental health (Cambridge, England), 8*, e7.

Kakuma, R., Minas, H., & Dal Poz, M. R. (2014). Strategies for strengthening human resources for mental health. In V. Patel, H. Minas, A. Cohen, & M. J. Prince (Eds.), *Global mental health: Principles and practice* (pp. 193–223). Oxford University Press.

Kanzler, K. E., Kilpela, L. S., Pugh, J., Garcini, L. M., Gaspard, C. S., Aikens, J., … Finley, E. P. (2021). Methodology for task-shifting evidence-based psychological treatments to non-licenced/lay health workers: protocol for a systematic review. *BMJ Open, 11*(2), Article e044012.

Keynejad, R. C., Dua, T., Barbui, C., & Thornicroft, G. (2018). WHO mental health gap action Programme (mhGAP) intervention guide: A systematic review of evidence from low and middle-income countries. *Evidence-Based Mental Health, 21*(1), 30–34. https://doi.org/10.1136/eb-2017-102750

Khan, M. N., Hamdani, S. U., Chiumento, A., Dawson, K., Bryant, R. A., Sijbrandij, M., … Rahman, A. (2019). Evaluating feasibility and acceptability of a group WHO trans-diagnostic intervention for women with common mental disorders in rural Pakistan: A cluster randomised controlled feasibility trial. *Epidemiology and Psychiatric Sciences, 28*(1), 77–87.

Kohrt, B. A., Jordans, M. J., Rai, S., Shrestha, P., Luitel, N. P., Ramaiya, M. K., et al. (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.

Landry, S. H., Assel, M. A., Carlo, M. S., Williams, J. M., Wu, W., & Montroy, J. J. (2019). The effect of the Preparing Pequeños small-group cognitive instruction program on academic and concurrent social and behavioral outcomes in young Spanish-speaking dual-language learners. *Journal of School Psychology, 73*, 1–20.

Laurenzi, C. A., Gordon, S., Skeen, S., Coetzee, B. J., Bishop, J., Chademana, E., & Tomlinson, M. (2020). The home visit communication skills inventory: Piloting a tool to measure community health worker fidelity to training in rural South Africa. *Research in Nursing & Health, 43*(1), 122–133.

Leeman, J., Birken, S. A., Powell, B. J., Rohweder, C., & Shea, C. M. (2017). Beyond "implementation strategies": Classifying the full range of strategies used in implementation science and practice. *Implementation Science : Iscus, 12*(1), 125. https://doi.org/10.1186/s13012-017-0657-x

Lehmann, U., & Sanders, D. (2007). Community health workers: What do we know about them? The state of the evidence on programmes, activities, costs an impact on health outcomes of using community health workers. World health organization. Available from: https://www.hrhresourcecenter.org/node/1587.html. (Accessed 13 February 2022).

Leocata, A. M., Kaiser, B. N., & Puffer, E. S. (2021). Flexible protocols and paused audio recorders: The limitations and possibilities for technologies of care in two global mental health interventions. *SSM. Mental health, 1,* Article 100036.

Lewis, C. C., Fischer, S., Weiner, B. J., Stanick, C., Kim, M., & Martinez, R. G. (2015). Outcomes for implementation science: An enhanced systematic review of instruments using evidence-based rating criteria. *Implementation Science : Iscus, 10,* 155.

Mastroleo, N. R., Mallett, K. A., Turrisi, R., & Ray, A. E. (2009). Psychometric properties of the Peer Proficiency Assessment (PEPA): A tool for evaluation of undergraduate peer counselors' motivational interviewing fidelity. *Addictive Behaviors, 34*(9), 717–722.

Michie, S., van Stralen, M. M., & West, R. (2011). The behaviour change wheel: A new method for characterising and designing behaviour change interventions. *Implementation Science : Iscus, 6,* 42. https://doi.org/10.1186/1748-5908-6-42

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., et al., PRISMA-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews, 4*(1).

Munodawafa, M., Lund, C., & Schneider, M. (2017). A process evaluation exploring the lay counsellor experience of delivering a task shared psycho-social intervention for perinatal depression in Khayelitsha, South Africa. *BMC Psychiatry, 17*(1), 236.

Murray, L. K., Dorsey, S., Bolton, P., Jordans, M. J., Rahman, A., Bass, J., et al. (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*(1), 30.

Nadkarni, A., Weobong, B., Weiss, H. A., McCambridge, J., Bhat, B., Katti, B., et al. (2017). Counselling for alcohol problems (CAP), a lay counsellor-delivered brief psychological treatment for harmful drinking in men, in primary care in India: A randomised controlled trial. *Lancet (London, England), 389*(10065), 186–195. https://doi.org/10.1016/S0140-6736(16)31590-2

Newnham, E. A., McBain, R. K., Hann, K., Akinsulure-Smith, A. M., Weisz, J., Lilienthal, G. M., et al. (2015). The Youth Readiness Intervention for war- affected youth. *Journal of Adolescent Health : Official Publication of the Society for Adolescent Medicine, 56*(6), 606–611. https://doi.org/10.1016/j.jadohealth.2015.01.020

O'Shea, O., McCormick, R., Bradley, J., & O'Neill, B. (2016). Fielity review: A scoping review of the methods used to evaluate treatment fidelity in behavioural change interventions. *Physical Therapy Reviews, 21*(3), 207–214.

Ottman, K. E., Kohrt, B. A., Pedersen, G. A., & Schafer, A. (2020). Use of role plays to assess therapist competency and its association with client outcomes in psychological interventions: A scoping review and competency research agenda. *Behaviour Research and Therapy, 130,* 1–3531. https://doi.org/10.1016/j.brat.2019.103531

Patel, V., Weiss, H. A., Chowdhary, N., Naik, S., Pednekar, S., Chatterjee, S., et al. (2011). Lay health worker led intervention for depressive and anxiety disorders in India: Impact on clinical and disability outcomes over 12 months. *British Journal of Psychiatry : The Journal of Mental Science, 199*(6), 459–466. https://doi.org/10.1192/bjp.bp.111.092155

Patel, V., Weiss, H. A., Chowdhary, N., Naik, S., Pednekar, S., Chatterjee, S., et al. (2010). Effectiveness of an intervention led by lay health counsellors for depressive and anxiety disorders in primary care in Goa, India (MANAS): A cluster randomised controlled trial. *Lancet (London, England), 376*(9758), 2086–2095. https://doi.org/10.1016/S0140- 6736(10)61508-5

Patel, V., Weobong, B., Weiss, H. A., Anand, A., Bhat, B., Katti, B., et al. (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. *Lancet (London, England), 389*(10065), 176–185. https://doi.org/10.1016/S0140-6736(16)31589-6

Perez, J., Lengnick-Hall, R. M., & Mittman, B. S. (2019). Core functions and forms of complex health interventions: A patient-centered medical home illustration. *Journal of General Internal Medicine, 34,* 1032–1038.

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., et al. (2011). Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. *Adm. Pol. Ment. Health, 38*(2), 65–76. https://doi.org/10.1007/s10488-010-0319-7

Puffer, E. S., Friis-Healy, E. A., Giusto, A., Stafford, S., & Ayuku, D. (2021). Development and implementation of a family therapy intervention in Kenya: A community-embedded lay provider model. *Global Social Welfare, 8,* 11–28.

Rahman, A., Akhtar, P., Hamdani, S. U., Atif, N., Nazir, H., Uddin, I., … Zafar, 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 (Cambridge England), 6,* e8.

Resnick, B., Inguito, P., Orwig, D., Yahiro, J. Y., Hawkes, W., & Werner, H. (2005). Treatment fidelity in behavior change research: A case example. *Nursing Research, 54*(2), 139–143.

Shahmalak, U., Blakemore, A., Waheed, M. W., & Waheed, W. (2019). The experiences of lay health workers trained in task-shifting psychological interventions: A qualitative systematic review. *International Journal of Mental Health Systems, 13,* 64. https://doi.org/10.1186/s13033-019-0320-9

Sileo, K. M., Miller, A. P., Wagman, J. A., & Kiene, S. M. (2021). Psychosocial interventions for reducing alcohol consumption in sub-saharan african settings: A systematic review and meta-analysis. *Addiction, 116*(3), 457–473.

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. *Annual Review of Clinical Psychology, 13,* 149–181.

Singla, D. R., Ratjen, C., Krishna, R. N., Fuhr, D. C., & Patel, V. (2020). Peer supervision for assuring the quality of non-specialist provider delivered psychological intervention: Lessons from a trial for perinatal depression in Goa, India. *Behaviour Research and Therapy, 130,* Article 103533.

Singla, D. R., Raviola, G., & Patel, V. (2018). Scaling up psychological treatments for common mental disorders: A call to action. *World Psychiatry, 17*(2), 226.

The President and Fellows of Harvard College. (2022). *Empower: Building the mental health workforce.* Harvard Medical School. https://mentalhealthforalllab.hms.harvard.edu/empower

Theobald, S., Brandes, N., Gyapong, M., El-Saharty, S., Proctor, E., Diaz, T., et al. (2018). Implementation research: New imperatives and opportunities in global health. *Lancet (London, England), 392*(10160), 2214–2228.

Tsai, A. C. (2017). Lay worker-administered behavioral treatments for psychological distress in resource-limited settings: Time to move from evidence to practice? *PLoS Medicine, 14*(8), Article e1002372. https://doi.org/10.1371/journal.pmed.1002372

Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. S. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology, 61*(4), 620–630. https://doi.org/10.1037//0022-006x.61.4.620

Wampold, B. (2015). How important are the common factors in psychotherapy? An Update. *World Psychiatry, 14*(3), 270–277.

Webb, C. A., Derubeis, R. J., & Barber, J. P. (2010). Therapist adherence/competence and treatment outcomes: A meta-analytics analysis. *Journal of Consulting and Clinical Psychology, 78*(2), 200–211.

World Health Organization. (2007). *Task shifting: Rational redistribution of tasks among health workforce teams: Global recommendations and guidelines.* Geneva: World Health Organization.

World Health Organization. (2008). *Task shifting: Rational redistribution of tasks among health workforce teams: Global recommendations and guidelines.* Geneva: WHO.

World Health Organization. (2022). EQUIP – ensuring quality in psychological support. https://www.who.int/teams/mental-health-and-substance-use/treatment-care/equip-ensuring-quality-in-psychological-support.

World Health Organization and United Nations High Commissioner for Refugees. (2015). *mhGAP Humanitarian Intervention Guide (mhGAP-HIG): Clinical management of mental, neurological and substance use conditions in humanitarian emergencies.* Geneva: WHO.