Administration and Policy in Mental Health and Mental Health Services Research (2022) 49:343–356
https://doi.org/10.1007/s10488-021-01167-x

ORIGINAL ARTICLE

# Knowledge and Attitudes Toward an Artificial Intelligence-Based Fidelity Measurement in Community Cognitive Behavioral Therapy Supervision

**Torrey A. Creed[1,6] · Patty B. Kuo[2] · Rebecca Oziel[1,6] · Danielle Reich[1,6] · Margaret Thomas[3] · Sydne O'Connor[1,6] · Zac E. Imel[2] · Tad Hirsch[3] · Shrikanth Narayanan[4] · David C. Atkins[5]**

Accepted: 7 September 2021 / Published online: 18 September 2021


## Abstract

To capitalize on investments in evidence-based practices, technology is needed to scale up fidelity assessment and supervision. Stakeholder feedback may facilitate adoption of such tools. This evaluation gathered stakeholder feedback and preferences to explore whether it would be fundamentally feasible or possible to implement an automated fidelity-scoring supervision tool in community mental health settings. A partially mixed, sequential research method design was used including focus group discussions with community mental health therapists (n = 18) and clinical leadership (n = 12) to explore typical supervision practices, followed by discussion of an automated fidelity feedback tool embedded in a cloud-based supervision platform. Interpretation of qualitative findings was enhanced through quantitative measures of participants' use of technology and perceptions of acceptability, appropriateness, and feasibility of the tool. Initial perceptions of acceptability, appropriateness, and feasibility of automated fidelity tools were positive and increased after introduction of an automated tool. Standard supervision was described as collaboratively guided and focused on clinical content, self-care, and documentation. Participants highlighted the tool's utility for supervision, training, and professional growth, but questioned its ability to evaluate rapport, cultural responsiveness, and non-verbal communication. Concerns were raised about privacy and the impact of low scores on therapist confidence. Desired features included intervention labeling and transparency about how scores related to session content. Opportunities for asynchronous, remote, and targeted supervision were particularly valued. Stakeholder feedback suggests that automated fidelity measurement could augment supervision practices. Future research should examine the relations among use of such supervision tools, clinician skill, and client outcomes.

**Keywords** Cognitive behavioral therapy · Fidelity · Supervision · Artificial intelligence · Machine learning · Technology · Community mental health

✉ Torrey A. Creed
tcreed@pennmedicine.upenn.edu

1 Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

2 Department of Educational Psychology, University of Utah, Salt Lake City, UT, USA

3 Northeastern University, Boston, MA, USA

4 Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA, USA

5 Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA

6 Penn Collaborative for CBT & Implementation Science, 3535 Market Street, Suite 3046, Philadelphia, PA 19104, USA

In 2017, over 46 million American adults had a diagnosed mental health disorder, and depression is expected to be the single greatest source of global disease burden by 2030 (Center for Behavioral Health Statistics & Quality, 2018). Effectiveness research has documented a variety of evidence-based psychotherapies (EBPs; e.g., cognitive-behavioral therapy; CBT) to treat these disorders (Hollon et al., 2020), and national, state, and regional mental health systems have implemented EBPs, compelling therapists to deliver them when clinically indicated (Creed et al., 2016a; McHugh & Barlow, 2010). Despite the investment of billions of dollars, policy mandates, and value-based incentives to disseminate EBPs into routine care settings, the quality of these sessions remains highly variable (Olfson & Marcus, 2010).

In the United States, therapists are expected to receive supervision through their clinical training process until they are licensed, and then to participate in continuing education for as long as they continue in clinical practice (Association of State and Provincial Psychology Boards, 2018). Supervisees and their licensed supervisors commonly meet weekly to review sessions that the supervisee conducted since their last meeting (Ronnestad & Skovholt, 1993). Typically, supervisees describe their sessions and the supervisor provides feedback about areas for growth, successes, and plans for upcoming sessions. Once a therapist earns a license, supervision usually ends and subsequent formal opportunities for skill development are rare (Tracey et al., 2014). Assessment of clinician skill is infrequent during training or while receiving supervision, and highly atypical after licensure, which may contribute to variability in quality of the treatment that is delivered (Olfson & Marcus, 2010). In fact, therapists' clinical skills may begin to erode over time (Schwalbe et al., 2014), which can have a deleterious impact on client symptom improvement (Goldberg et al., 2016).

Key components of effective supervision include the provision of accurate, consistent, actionable feedback based on a therapist's in-session behavior (Newman & Kaplan, 2016), and participation in this kind of supervision can maintain and improve a therapist's clinical skills (Creed et al., 2016a, 2021; Schwalbe, et al., 2014; Tracey et al., 2014). However, provision of detailed feedback or fidelity assessment is time consuming, expensive, and reliant on access to supervisors or expert consultants; it is a non-starter in the vast majority of real-world settings (Stirman et al., 2018). Accordingly, mental health services researchers have developed a variety of alternative measures, including patterns of utilization, therapist self-reports of adherence, client rated measures of satisfaction, or measures of clinical outcomes (England & Butler, 2015). However, these are proxies of intervention quality, distal to the content of the clinical encounter, may be subject to self-observation bias, and are rarely available to therapists in an immediate, clinically actionable form. Without performance-based feedback or quality monitoring, the impact of costly dissemination efforts is often not sustained (Proctor et al., 2011; Schwalbe et al., 2014); conversely, when therapists are provided with regular feedback, patient outcomes improve (Anker et al., 2009; Lambert et al., 2001).

While discussion of a therapist's report of a session can guide supervision feedback, more specific and detailed information about their session behavior can be crucial for improving clinical skills (Tracey, et al., 2014). Session recordings can provide more objective information about what occurred in session, including details that a therapist may have missed, without significantly impacting the therapeutic processes (Briggie et al., 2016; Brown et al., 2013). In fact, session recordings may provide more accurate information about therapeutic interactions than a therapist's self-report (Mehr et al., 2010; Waltman et al., 2016), and therapists' evaluation of their own CBT skills has been found to be unrelated to skills demonstrated in session (Creed et al, 2016b, 2021). Recordings may therefore facilitate the delivery of specific, concrete feedback in supervision (Waltman et al., 2016). Given the paucity of opportunities for skill development post-licensure, review of recordings may also provide more seasoned therapists with the opportunity for skill development.

Although the integration of session recording and review into supervision holds great potential, these strategies are very time and resource-intensive and may not be scalable for the large system-level implementation of evidence-based practices that have been undertaken (e.g., Clark, 2011; Karlin et al., 2012; McHugh & Barlow, 2010). Proctor et al. concluded that, "The foremost challenge [to disseminating psychosocial EBPs] may be *measuring implementation fidelity quickly and efficiently*" (p. 70; italics added; Proctor et al., 2011). To ensure that clients are receiving EBPs as they were intended to be delivered, understand reasons for potential differences in effectiveness or engagement, and capitalize on the significant investment that health systems have made in EBPs, technology is needed to scale up fidelity assessment "quickly and efficiently," facilitating specific and data-driven supervision.

Advances in artificial intelligence (AI), including natural language processing and machine learning, offer methods for recognizing patterns in spoken language that predict indicators of fidelity in therapy session recordings, without the rate-limiting factors associated with reliance on humans to review sessions. For example, research teams have developed and evaluated a system for automated fidelity ratings for an EBP for substance abuse—motivational interviewing (MI)—called the Counselor-Observer Ratings Expert for MI (CORE-MI; Hirsch et al., 2018; Imel et al., 2019; Xiao et al., 2015). The system uses speech signal processing to review recordings of a therapy session, then generates a report of how well the therapist met MI fidelity standards (Hirsch et al., 2018). CORE-MI embeds this MI fidelity report in a performance-based feedback system that provides report-card like feedback on MI skills with data visualization features (Kuo et al., under review). This interactive tool allows therapists to record, review, and comment on videos of their therapy sessions, including timestamps that allow the user to skip to a flagged moment in the video when they click on the comment box. The tool also provides a searchable transcript of the session, which can be used to identify key moments for supervision. Preliminary results regarding the use of this type of interactive, recording-based platform and fidelity evaluation system in daily clinical practice suggest that therapists found that CORE-MI facilitated personal reflection about their clinician skills and better engagement in clinical supervision (Kuo et al., under review). Foundational

research has now extended the use of machine learning technology for behavioral coding beyond MI to the evaluation of CBT. In a series of papers, our team has used a large corpus of human ratings of the Cognitive Therapy Rating Scale (CTRS; Creed et al., 2021; Young & Beck, 1980) to train machine learning models that can automatically generate CTRS scores for a session recording (Gibson et al., 2019; Flemotomos et al. 2018).

There is tremendous promise in the use of machine learning technology to support the evaluation of psychotherapy in the community. However, new innovations from university-based research labs do not automatically—nor often—translate into more effective practice in the community. Relative to research samples, community settings may have different client populations, organizational climates, staff attitudes, clinical workflows, and administrative structures. For example, in an evaluation study of CORE-MI, more experienced therapists tended to share more skepticism about the accuracy of an automated report (Hirsch et al., 2018), and a subsequent study of CORE-MI users found that more seasoned therapists saw more value for trainees than for their own clinical practice (Kuo et al., under review). Interventions and tools that are developed in partnership with stakeholders in the contexts in which they will be used may maximize impact and uptake (Jull et al., 2017; McLean & Tucker, 2013). The integration of stakeholder feedback regarding feasibility and applicability to their settings has the potential to facilitate rapid adoption of new tools (Proctor et al., 2009; Weiner et al., 2017). A parallel philosophy exists within technology development known as user-centered design, "… an approach that puts human needs, capabilities, and behavior first, then designs to accommodate those needs, capabilities, and ways of behaving" (Norman, 2002, p. 8). Both implementation research and user-centered design begin with needs assessment and rich observation of the clinical context in which a treatment or technology will be deployed, and development proceeds through an iterative process of end-user feedback and refinement. Ultimately, successful treatments and technologies are developed not simply 'for' the real world but in the real world.

The goal of this study was to understand how community mental health therapists and clinical leadership perceive AI-based automated evaluation and assessment as a supervision tool to provide feedback on CBT fidelity. Specifically, we examined community mental health providers' perceptions of the strengths and needs of typical CBT supervision, as well as their perceptions (e.g., fit, feasibility, acceptability, concerns, wishes) related to a proposed adaptation of CORE-MI to support CBT supervision and fidelity rating. In addition, we gathered information about community mental health providers' capacity, knowledge, and experience with the types of technology that comprise an automated fidelity-rating system in order to evaluate the needs of community mental health providers, the extent to which their context might have the prerequisite tools to deploy an AI-based supervision tool, and the extent to which they were familiar with the underlying technology.

## Methods

### Overview

This mixed-methods study included semi-structured focus group discussions and online survey data to evaluate provider attitudes and practices related to standard supervision and the adaptation of CORE-MI to rate CBT fidelity (a proposed CORE-CBT) as a tool for supervision. We used a partially mixed, sequential research method design for significance-enhancement in order to maximize our understanding of therapist experiences with, and perceptions of, AI in clinical contexts. We primarily used qualitative methodologies to gain a more nuanced, rich understanding of how community mental health therapists and clinical leadership engage in supervision and would apply the proposed CORE-CBT to their supervision practices. We sought to enhance the interpretation of our qualitative findings by using quantitative measures of (a) tparticipants' current use of AI, and (b) perceptions of applications of AI in clinical contexts, to complement and provide more context to participant narratives (Leech & Onwuegbuzie, 2010; Onwuegbuzie & Leech, 2004).

### Study Context

This study was conducted in the context of a High-Priority, Short Term Project (R56) funded by the National Institute of Mental Health focused on development of a technology to assist in training and supervision by providing therapist CBT fidelity ratings for individual psychotherapy sessions. Participants were drawn from organizational partners in the University of Pennsylvania's Beck Community Initiative (Penn BCI), which is a public-academic partnership including the City of Philadelphia's Department of Behavioral Health and Intellectual disAbility Services (DBHIDS), and Philadelphia's community mental health care providers. (See Creed et al., 2014 for a description of the Penn BCI). Focus group participants were recruited in Fall 2019 from adult outpatient community mental health provider organizations within the Penn BCI. Specifically, participants represented three programs—two outpatient therapy programs focused on general mental health care that had been part of the Penn BCI for 4 and 10 years, respectively, and one outpatient therapy program focused on substance abuse services that had been part of the BCI for 3 years. These organizations were selected based on their active participation status with

the BCI, their level of care, and their interest in participation. Focus group participants were employed by these organizations but were not limited to therapists who had received training with the Penn BCI. Informed consent was obtained from all individual participants included in the study, and participants were each compensated $25 for their participation. All study procedures were reviewed and approved by the first author's Institutional Review Board, and the study was performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments.

## Participants

Study participants included 30 unique mental health care therapists ($n = 18$), supervisors ($n = 8$), and other clinical leadership (e.g., clinical directors; $n = 4$) from organizations that had participated in a training program to implement CBT in their clinical services. Most were female (75.86%), held a masters degree or higher (66% of therapists, 100% of clinical leadership), and worked in their position for an average of 6.18 years, although this ranged from 5 months to 20 years. To recruit a criterion sample of participants representing all implementing roles, the research team collaborated with the organization. Specifically, a member of the clinical leadership at the agency invited all therapists (regardless of whether they had participated in the Penn BCI program) to participate in voluntary focus groups. The supervisors and clinical directors of those staff were then identified by the agency contact and also invited to participate, yielding 18 therapists and 12 clinical leadership participants, all of whom contributed to both the quantitative and qualitative data.

## Procedure

### Focus Groups

Focus groups were held on-site at each of the three agencies. Participants were split into two simultaneously held groups per site based on their job role (e.g., therapists, supervisors and other clinical leadership), with 5–7 participants per group. Groups were separated because supervisors and therapists might, by nature of their roles, have different priorities or opinions about current and desired supervision strategies. Each group was also attended by a focus group facilitator who had been trained in focus group facilitation, and a note-taker from the study team. To build rapport, each research team member introduced themselves briefly at the beginning of the discussion, described their role on the team, and shared how the team would utilize general themes from the focus group to evaluate and refine the tool's development. Participants also introduced themselves and their role at the organization prior to the beginning of the discussion.

Each facilitator followed a semi-structured interview format to lead participants through a three-step discussion (see Online Appendix 1 for interview protocols) that lasted for 60 min. First, participants were asked a series of questions about their experiences and preferences related to supervision and training in CBT. Next, participants viewed a 7-min video that demonstrated CORE-MI as an illustration of the technology, with additional information about proposed CBT adaptations. Finally, participants were asked a series of questions about their opinions and preferences about an adaptation of CORE-MI for CBT. All focus groups were video-recorded with a second audio-recording to supplement transcription. The note-taker also took synchronous notes indicating who was speaking to support accurate transcription. Discussions were professionally transcribed verbatim and supplemented by the note-taker's notations.

### Surveys

All participants completed a series of online surveys in the 2 weeks prior to the focus groups. The measures of acceptability, appropriateness and feasibility of the CORE-CBT tool were re-administered immediately after focus groups were held in order to provide more context for participant's qualitative responses before and after seeing the CORE-MI demonstration. More specifically, we were interested in whether participant perceptions of applications of CORE-CBT may have changed after seeing a demonstration of CORE-MI and discussion among fellow participants. Consequently, we obtained pre-post measures of acceptability, appropriateness, and feasibility to augment how participants' qualitative experiences may have shifted.

**Access and Use of Technology** Participants responded to 9 questions about their current access and use of devices on which the CORE-CBT technology could be accessed in order to assess the extent to which these community mental health contexts currently had access to the prerequisite technology necessary to implement an AI-based supervision tool.

**Recognition of AI** The Recognition of AI Survey (Zhang & Dafoe, 2019) lists 10 examples of AI that are frequently used by the public, in order to evaluate participants' basic familiarity with AI. Respondents are asked, "In your opinion, which of the following technologies, if any, uses artificial intelligence (AI)? Select all that apply." Other than a set of norms published in 2019 based on a survey of 2000 American adults, there are no formal scoring procedures or psychometric data for the Recognition of AI survey. As

such, the data collected has been used for descriptive purposes only.

**Acceptability, Appropriateness, and Feasibility** The constructs of acceptability, appropriateness, and feasibility are often used as leading indicators of implementation outcome and are conceptually distinct (Powell et al., 2017). Weiner et al., (2017) developed brief measures of each construct. The Acceptability of Intervention Measure (AIM) measures the perception that a given practice or innovation is agreeable, palatable or satisfactory. The Intervention Appropriateness Measure (IAM) examines stakeholders' sense of perceived fit, relevance or compatibility of the practice or innovation for their practice setting, clients, or treatment population. The Feasibility of Intervention Measure (FIM) evaluates perceptions of how well the practice or innovation could be successfully used within the stakeholders' setting. Each measure includes 4 statements, and respondents rate their agreement with each statement on a 5-point scale ranging from 1 (completely disagree) to 5 (completely agree). Ratings are totaled, and higher scores indicate greater acceptability, appropriateness, or feasibility. Test–retest reliability coefficients for each measure ranged from 0.73 to 0.88, and regression analysis indicated each was sensitive to change in both directions (Weiner et al., 2017).

## Data Analysis

### Focus Group Data Analysis

Transcripts were analyzed by a team of three coders and one auditor using a multi-step process and managed using Dedoose (2020). The first coder was a master's level therapist with clinical and research experience in CBT. The second coder completed a master's degree in design, with a specific focus on user-centered design; she was familiar with application of AI in clinical contexts and CBT in group therapy. The third coder had a bachelor's degree in psychology, with undergraduate and post-graduate research and clinical experience. The auditor has a doctorate in clinical psychology, and was a research and clinical expert in CBT; she supervised the team of coders and has led research related to the role of AI in fidelity monitoring and supervision.

Given the stated interest in understanding stakeholders' experiences and perceptions of application of AI to clinical work, research questions were developed and data were coded using a phenomenological approach (Groenewald, 2004). Phenomenological design was developed as a method of describing specific psychological and social phenomena from the perspective of individuals affected by phenomena of interest, and understanding their internal experiences. Furthermore, phenomenological approaches allow for close examination of contextual factors that contribute to how individuals experience specific phenomena (Groenewald, 2004). Consequently, a phenomenological design was an ideal method of exploring and understanding therapist perceptions of AI methodologies in fostering the development of CBT skills, and how these perceptions may be informed by their clinical experiences.

Transcripts were analyzed using an inductive coding approach that started with open coding, which involved developing an exhaustive familiarity with the source data (Charmaz, 2014; Glaser, 1978). This process involved repeated independent readings of the transcripts over several weeks, line-by-line coding of one randomly selected transcript, and development of an initial codebook from these codes. Next, the coders engaged in axial coding and developed higher level themes by comparing and examining shared commonalities between codes. The coders then applied the codebook to each transcript and independently coded each transcript line-by-line. Coders constantly compared codes, and resolved coding disagreements with the assistance of the auditor. Prior to resolving coding disagreements, an interrater reliability analysis using the Kappa statistic was performed to determine consistency among raters (all Kappa $\geq 0.78$, $p < 0.01$). The coding team revisited and adjusted themes and subthemes in the codebook throughout the coding process until saturation was reached. Further, coders discussed how values, experiences, and biases informed their codes throughout the coding process.

### Survey Data Analysis

Survey data were summarized using descriptive statistics. The mean and standard deviation are reported for continuous measures, and frequency and percentages are reported for categorical variables. A Wilcoxon signed rank test for paired samples was employed to compare means for measures that were repeated after the focus groups; this non-parametric test was selected because the non-normality of the data violated the assumptions of parametric tests.

# Results

## Focus Groups

Analysis of the focus group responses revealed 5 major themes: typical supervision practices, as well as the utility, abilities, challenges or concerns, and desired features of an automated fidelity measurement system, using the existing CORE-MI tool and proposed CBT adaptations as an example. Summary of key themes along with illustrative participant quotes are provided in Table 1.

**Table 1** Qualitative themes, codes, and illustrative quotes from clinical leadership (CL) and therapists (T)

| Theme | Codes | Illustrative quotes |
|---|---|---|
| Typical supervision practices | (a) Focus of supervision<br>(b) Driver of supervision<br>(c) Supervision tool needs | (a1) "For me personally, it's going over any specific clinical questions, going through their caseload, and teaching them—whether it's CBT, whether it's DBT skills, whether it's play therapy—teaching them clinical skills that they can use with their clients." (CL)<br>(a2) "I'm still learning, so I'll go to my supervisor and be like, 'I don't know what to do here,' and then she'll give me like techniques to use or she'll say, "Hey, why don't you try this? Go look at it and then come back with questions." (T)<br>(b1) "For individual supervision, [therapists] prepare an agenda and come in, and we talk about that." (CL)<br>(b2) "I usually decide what to talk about. I go by whether I have anyone with significant suicidality or homicidality." (T)<br>(c1) "I think my biggest need is to find a, some sort of organized way of having forms. I have 800 million different forms scattered around. There are just so many pieces of paper! To have something where you can have at our fingertips—things we can access right away." (CL)<br>(c2) "Just getting supervision regularly would be really great." (T) |
| Utility | (a) Utility for supervision<br>(b) Utility for training<br>(c) Utility for professional growth | a1) "I think it would give us the opportunity to see things that we've never touched on in supervision because you've never had access to it." (CL)<br>(a2) "Being able to have the notation in there and archiving it means that I can have that information for a longer period of time, and it allows me to have that discourse when I'm between supervisions, so that I can still grow and then can talk about it in supervision." (T)<br>(b1) "I can see this enhancing all kinds of things like Socratic questioning … guided discovery and everything like that. Because when you're first learning CBT, you do a lot of talking. Too much talking probably." (CL)<br>(b2) "I think it would be helpful for someone who's brand new to get some sort of baseline—and then being able to… really build skills and focus on your weaknesses. That'd be so helpful—ease some anxiety." (T)<br>(c1) "I'm competitive. I feel like, 'How do I get my score up?'" (CL)<br>(c2) "I think it's a good thing to hold ourselves accountable to asking those feedback questions." (T) |
| Nuance | (a) Non-verbal content<br>(b) Rapport<br>(c) Cultural diversity | (a1) "Our [clients] tend to be more highly emotionally reactive. So sometimes the intervention is not reacting and just listening, which may not appear empathetic. That'd be a visual thing." (CL)<br>(a2) "What about body language? Sometimes the look on my face makes a difference in how to take what I say." (T)<br>(b1) "The empathy score does make me nervous. That's the only part that I'm like, "How does the computer know if we're being empathic?" (CL)<br>(b2) "Every relationship is different. How will it know if my client thinks I get them?" (T)<br>(c1) "Can (the tool) understand things like different parenting practices in different cultures?" (CL)<br>(c2) "Yeah, and what is empathy across cultures?" (T) |
| Challenges or concerns | (a) Organizational restrictions<br>(b) Privacy<br>(c) Self-confidence | (a1) "I don't know if (agency name) would budge on the idea of video, because of how we had to fight to get [audio] recordings." (CL)<br>(a2) "I'm worried about our computer's capabilities of handling that." (T)<br>(b1) "I'm wondering if people might be less likely to talk about things if they're being recorded." (CL)<br>(b2) "I could see people getting access to it who are not the [client] … and to me that's a huge confidentiality risk." (T)<br>(c1) "I can see how these scores would help someone. If the scores are going up, they might be more confident.'" (CL)<br>(c2) "I feel like it would take some getting used to. There's room to be really critical of yourself." (T) |

**Table 1** (continued)

| Theme | Codes | Illustrative quotes |
|---|---|---|
| Desired features | (a) Valued features<br>(b) Wishes | (a1) "Let's say there's an intern… who doesn't have time to meet with their supervisor. The supervisor can go on that platform, give really helpful, targeted supervision feedback, and they never even see each other. That's amazing, because I feel like that's one of the biggest problems—having time for the supervisee and supervisor to be able to meet." (CL)<br>(a2) "The searchable transcript would make it so easy to find the parts I need to talk about in supervision." (T)<br>(b1) "When we would get [feedback at] the 3-month and the 6-month [time point in BCI training], they would tell you, this is why you got that score, but then here's how you can get a little higher next time. So if it can do that, that would be awesome." (CL)<br>(b2) "A feature I would want to see is if the computer is able to identify the intervention that was used." (T) |

## Typical Supervision Practices

Throughout the conversations about supervision, and regardless of job role, participants emphasized the importance of daily in-person communication during regular staff meetings, one-on-one supervision, other formal or informal meetings, or via email and phone. Participants reported feeling as though they were "constantly talking about" clinical interventions and their implementation with their supervisors. Supervisors confirmed that regular discussion was intended to help therapists improve their CBT skills, make sense of new procedures, and monitor CBT use over time. Through these interpersonal interactions, supervisors disseminated and synthesized information, mediated between strategy and day-to-day tasks, and supported integration of EBPs into routine daily practice.

### Focus of Supervision

Focus of discussion in supervision varied, as did the amount of structure in supervision meetings. Generally, supervisors and therapists reported that supervision meetings included case updates and conceptualizations, as well as administrative issues and paperwork. There was also time allocated to training/teaching the therapist on particular skills, but this was more strongly emphasized when supervising interns or newer employees. A subset of respondents also reported a focus on self-care and the therapist's well-being.

### Driver of Supervision Content

Overall, participants reported that the topical focus of supervision meetings was decided collaboratively between supervisor and therapist. For participants who reported that supervision topics were expressly identified and then discussed, the responsibility for choosing topics might be either the supervisor or supervisee in a given dyad. High-priority topics tended to take precedence in driving the content as well, such as looming audit requirements or clinical cases with emergent needs.

### Supervision Tool Needs

In regard to current supervision practices that were valued, many participants reported using, and liking the use of, audio recordings. Participants also valued being able to consult with colleagues for advice and input, either impromptu or in group supervision. A subset of therapists noted a lack of consistency in their current supervision, such as not having a consistent supervisor or a consistent supervisory procedure.

## Utility

Participant reactions to CORE-MI and proposed adaptations for CBT-rating were overall positive. They particularly valued the aspects of CORE-MI that they perceived to measure more concrete constructs, like measurement of talk turns and creation of transcripts. Evaluation of constructs they perceived to be more subjective, like empathy or humor, was received with more anxiety and uncertainty. Participants primarily identified the utility of CORE-MI, and adaptations for CBT, in relation to training and supervision, and facilitation of professional growth.

### Utility for Training and Supervision

The ability to mark and review time-linked notes in sessions, and search transcripts, were particularly well received across groups and seen as useful for supervision. Many participants valued the feedback and scoring aspect of the tool, including the speed at which scores were available to users. In particular, participants highlighted the immediacy of feedback as useful for skill building during the learning or training process**.** Participants valued the ability to access both the

recording and searchable transcript to facilitate targeted feedback in supervision.

Many participants also agreed that a CBT-adaptation of CORE-MI would be helpful in managing time for CBT supervision (e.g., finding things quickly in the searchable transcript, reviewing before supervision). This capability was noted to be of particular value, given the agreement across groups that there generally was not enough available time for supervision. Many noted that the ability to provide commentary on specific moments in session would allow asynchronous supervision, broadening their overall supervision opportunities.

#### Utility for Professional Growth

Participants saw utility for automated rating in understanding baseline skill, tracking overall improvement, reinforcing strong skills, and identifying skills that require additional growth. Several participants also perceived the automated rating to be more objective than human ratings, which they valued in being able to assess their skills.

There were also several mentions of CORE-MI (or its adaptation for CBT) as a method of seeing which interventions worked in a specific session in order to focus on continuing to build on those successes with a client. Participants also shared how CORE-MI, and similar tools tailored for CBT, could facilitate greater accountability for therapists and encourage critical reflection on continued growth.

### Nuance

Participants raised questions related to whether AI could rate skills in a way that was sensitive to more nuanced facets of human interactions. Specifically, questions were raised about the ability of CORE-MI or a CBT adaptation to accurately rate skills like empathy, appropriate silences, tone of voice, rapport development, and responsiveness to cultural diversity.

#### Nonverbal Content and Rapport

The video feature was perceived as valuable to assess for nonverbal cues, both for accurate scoring by the tool, and for reference in supervision to provide clinically useful info. Participants expressed that AI tools could facilitate greater awareness of therapist-client dynamics, particularly those that they may be unaware of. However, among agencies that currently record sessions, recordings are almost exclusively audio-based, and while CORE-MI captures video content, the machine learning algorithms that yield fidelity scores rely exclusively on audio input. Consequently, participants were also concerned that CORE-MI and other similarly based AI tools may not holistically assess nonverbal cues. Several participants who expressed concern around AI based scores did note that explanations of how the scores were generated would increase their confidence in the scores.

#### Cultural Diversity

Some participants also expressed concerns about how CORE-MI and other AI-based tools pick up on variations in cultural norms such as discussions about culturally-appropriate behavior, differences in verbal/nonverbal expressions, and pronunciation. Other participants shared concerns of how CORE-MI and other AI-based tools may not capture important culture-specific information related to empathy or cultural practices.

### Challenges or Concerns

Supervisors and therapists reported concerns surrounding CORE-MI that centered on confidentiality and the privacy implications of clients potentially accessing their clinical records. Participants also noted that organizational restrictions may impede full implementation of software similar to CORE-MI. Further, participants described concerns about how implementation of software like CORE-MI could impact confidence in their clinical skills.

#### Privacy

Supervisors and therapists worried that if clients were given access to session recordings, transcripts, ratings, or comments about their sessions, the access could negatively impact the therapeutic relationship or their progress in therapy by hindering client disclosures. Some participants also expressed concerns that if clients were given access to the tool or ratings, they would inappropriately share information about their sessions with others and cause harm to themselves. Participants indicated that that clinical information should therefore not be accessible from outside the agency for clients or clinical staff, for risk of content being accessed inappropriately by a third party (like a spouse). Therapists (though not supervisors) also voiced concerns about their own privacy, perceiving it to be personally intrusive if video of their sessions were available to clients. Finally, participants expressed concerns about how information would be stored and secured, and whether it would be HIPAA compliant.

#### Organizational Restrictions

Participants also expressed concern about their organizations' ability to meet the technological requirements to host an automated fidelity-scoring tool, noting that resources in community mental health are limited and that organizations

are rarely able to buy or support computers, software or voice recorders. There was also consistent concern about agencies' policies related to video recording of sessions, reporting that this was against many of the agencies policies. Most also had concerns about their own skills or available time to manage video recording if that were required.

### Self-confidence

Several participants noted that low scores could cause therapists to doubt themselves or their clinical abilities. Conversely, others mentioned that receiving scores and feedback could help ease anxiety about performance or meeting skill expectations during training, because it would give them a sense of which skills they needed to strengthen prior to evaluation.

## Desired Features

Discussion of specific scores for CBT feedback was centered around the Cognitive Therapy Rating Scale (CTRS; Young & Beck, 1980). The CTRS, an 11-item expert-rated assessment of therapist skill, is the most common and widely-used measure of CBT fidelity (Goldberg et al., 2020; Muse & McManus, 2013). The CTRS is also the measure used to evaluate CBT competence in the Penn BCI, the program through which participants were previously trained in CBT (Creed et al., 2016a, 2021). Participants indicated that anchoring a CBT adaptation of CORE-MI in items similar to the CTRS made intuitive sense and was appealing to them.

More broadly, most participants agreed that a CBT-adaptation of CORE-MI would be helpful in managing time for CBT supervision (e.g., finding things quickly in the searchable transcript, reviewing before supervision, asynchronous supervision). This capability was noted to be of particular value, given the agreement across groups that there generally was not enough available time for supervision. Several specific features of the existing CORE-MI tool were also highlighted frequently across groups. The transcript search function was particularly well received and described as useful. Many participants valued the feedback and scoring aspect of the tool, including the speed at which scores were available to users. Almost all agreed that ability to timestamp and comment on moments in the session were of particular value.

A number of participants also suggested that information that provided context for their scores would also be of use, including information about which parts of the session were scored on a given item or ways in which they could improve their scores. Several participants also mentioned that they would like the system to identify specific interventions that were used, with a focus on those that were used well. Finally, clinical leadership expressed a desire for a better way to store and condense files and paperwork (e.g. electronically).

## Survey Results

Table 2 reports the summary of survey responses about technology access and use. Most therapists and clinical leadership reported having access to a desktop computer at work (77.78%; 83.33% respectively) that they use daily (94.44%; 75.00% respectively). Therapist and clinical leadership access to other technology at work was quite limited inside or outside of session, and most frequently used laptops in session (16.67%; 41.67% respectively) or smartphones outside of session (38.89%; 50.00% respectively). Technology use outside of work was more frequent and varied, with therapists and clinical leadership most frequently reporting

**Table 2** Device access and use among therapists (n = 18) and clinical leadership (n = 12)

| Device | Access at work n (%) | | | Frequency of use at work n (%) | | | Frequency of use outside of work n (%) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Access | Therapists | Leadership | Days per week | Therapists | Leadership | Days per week | Therapists | Leadership |
| Smartphone | None | 10 (55.56) | 2 (16.67) | 0–2 | 14 (77.78) | 3 (25.00) | 0–2 | 2 (11.11) | 1 (8.33) |
| | In session | 1 (5.56) | 4 (33.33) | 3–4 | 1 (5.56) | 4 (33.33) | 3–4 | 5 (5.56) | 0 (0.00) |
| | Out of session | 7 (38.89) | 6 (50.00) | 5 or more | 3 (16.67) | 5 (41.67) | 5 or more | 15 (83.33) | 11 (91.67) |
| Desktop computer | None | 2 (11.11) | 1 (8.33) | 0–2 | 1 (5.56) | 2 (16.67) | 0–2 | 15 (83.33) | 7 (58.33) |
| | In session | 14 (77.78) | 10 (83.33) | 3–4 | 0 (0.00) | 1 (8.33) | 3–4 | 2 (11.11) | 1 (8.33) |
| | Out of session | 2 (11.11) | 1 (8.33) | 5 or more | 17 (94.44) | 9 (75.00) | 5 or more | 1 (5.56) | 4 (33.33) |
| Laptop | None | 14 (77.78) | 4 (33.33) | 0–2 | 15 (83.33) | 7 (58.33) | 0–2 | 9 (50.00) | 3 (25.00) |
| | In session | 3 (16.67) | 5 (41.67) | 3–4 | 1 (5.56) | 1 (8.33) | 3–4 | 4 (22.22) | 4 (33.33) |
| | Out of session | 1 (5.56) | 3 (25.00) | 5 or more | 2 (11.11) | 4 (33.33) | 5 or more | 5 (27.78) | 5 (41.67) |
| Tablet | None | 16 (88.89) | 9 (75.00) | 0–2 | 15 (83.33) | 11 (91.67) | 0–2 | 12 (66.67) | 6 (50.00) |
| | In session | 1 (5.56) | 2 (16.67) | 3–4 | 3 (16.67) | 1 (8.33) | 3–4 | 4 (22.22) | 3 (25.00) |
| | Out of session | 1 (5.56) | 1 (8.33) | 5 or more | 0 (0.00) | 0 (0.00) | 5 or more | 2 (11.11) | 3 (25.00) |

daily use of a smartphone (83.33%; 91.67% respectively). Most therapists (n = 11, 61.11%) and half of clinical leadership (n = 6, 50%) indicated that their present workday was positively impacted by the use of a computer or computer-based technology, but half of therapists (n = 9, 50%) and most of the clinical leadership (n = 9, 75.00%) thought that they might need additional training if computer-based technology became more central to their workday.

In regard to recording sessions, most therapists indicated that they currently record sessions using a digital audio recorder (n = 12, 66.67%) or tape recorder (n = 6, 33.33%), and only 1 (5.56%) therapist indicated using video recordings. Almost all indicated that they would definitely (n = 13, 72.22%) or maybe (n = 3, 16.67%) be comfortable being recorded, and that their clients would definitely (n = 6, 33.33%) or maybe (n = 10, 55.56%) be comfortable being recorded. Clinical leadership indicated that they typically review these recordings before (n = 7, 58.33%), during (n = 8, 66.67%), and after (n = 6, 50%) supervision sessions.

Table 3 compared participant knowledge to American norms about what the public considers AI. Therapists overall were able to identify different examples of technology as AI just over half of the time (56.6%), and clinical leadership overall identified AI correctly 68.33% of the time, in comparison to American norms of 46.09%.

Raw comparisons of participants' perceptions of the acceptability, appropriateness, and feasibility of the use of AI to measure CBT fidelity suggested that therapist perceptions were somewhat higher than those of clinical leadership both prior to the focus groups and after focus group discussion of the CORE-MI example and proposed CBT adaptations (see Table 4). Although norms are not yet available for the AIM, IAM, and FIM (Weiner et al., 2017), mean ratings for both groups were higher than the measures' midpoints both before and after their exposure to CORE-MI. A Wilcoxon Signed-Ranks test indicated that therapists' scores of appropriateness were significantly higher after discussion of CORE-MI and proposed CBT adaptations, and clinical leaderships' scores of acceptability, appropriateness, and feasibility were all significantly higher positive after the focus group discussion.

**Table 3** Percentage of respondents who identified technology as AI

| Technology | Therapists %(n) | Leadership %(n) | Norms[a] % |
|---|---|---|---|
| Virtual assistants (e.g., Siri, Google Assistant, Amazon Alexa) | 100.00 (n = 18) | 91.67 (n = 11) | 62.87 |
| Smart speakers (e.g., Amazon Echo, Google Home, Apple Homepod) | 72.22 (n = 13) | 75.00(n = 9) | 55.46 |
| Facebook photo tagging | 44.44 (n = 8) | 58.33(n = 7) | 36.16 |
| Google search | 61.11 (n = 11) | 50.00(n = 6) | 35.59 |
| Recommendations for Netflix movies or Amazon ebooks | 38.89 (n = 7) | 66.67(n = 8) | 27.73 |
| Google translate | 38.89 (n = 7) | 41.67(n = 5) | 29.49 |
| Driverless cars and trucks | 61.11 (n = 11) | 83.33(n = 10) | 56.38 |
| Social robots that can interact with humans | 66.67 (n = 12) | 83.33(n = 10) | 63.63 |
| Industrial robots used in manufacturing | 33.33 (n = 6) | 66.67(n = 8) | 40.11 |
| Drones that do not require a human controller | 50.00 (n = 9) | 66.67(n = 8) | 53.48 |
| Overall mean | 56.60 | 68.33 | 46.09 |

[a]Zhang & Dafoe (2019)

**Table 4** Self-reports of acceptability, appropriateness, and feasibility before and after focus groups

| | Therapists | | Significance | Leadership | | Significance |
|---|---|---|---|---|---|---|
| | Before focus groups Median; m (SD) | After focus groups Median; m (SD) | | Before focus groups Median; m (SD) | After focus groups Median; m (SD) | |
| Acceptability | 4; 3.84 (0.71) | 4; 4.28(0.56) | Z = − 2.73, p = .06 | 3.88; 3.77 (0.63) | 4; 4.19 (0.62) | Z = − 2.11, p = .04 |
| Appropriateness | 4; 3.72 (0.74) | 4; 4.07 (0.73) | Z = − 1.98, p = .048 | 3; 3.33 (0.78) | 4; 4.04 (0.72) | Z = − 2.73, p = .006 |
| Feasibility | 4; 3.99 (0.62) | 4; 4.06 (0.59) | Z = − 0.41, p = 0.68 | 3.62; 3.60 (0.54) | 3.88; 4.10 (0.69) | Z = − 2.53, p = .01 |

## Discussion

This mixed methods evaluation of community mental health therapists, supervisors, and other leadership sought to explore whether it would be fundamentally feasible or possible to implement an automated fidelity-scoring tool in community mental health care settings. Would attitudes, policies, or practices support its use? To explore these questions and inform future design in ways that may facilitate adoption of such tools (Proctor et al., 2009; Weiner et al., 2017), this study examined stakeholder feedback about standard supervision practices, access to technology, and reactions to a proposed automated CBT fidelity tool. In sum, feedback suggested that community providers in this large public mental health system perceive an AI-based supervision platform for CBT to be acceptable, appropriate, and feasible, and that they have the infrastructure in place to use such a system. While perceptions of the tool were overall positive, participants raised questions and concerns that should guide future tool development and strategies for its implementation. Findings from this study set the stage for future research to refine and implement technology-based supervision and evaluation tools, which in turn may have implications for improving access to high-quality delivery of EBPs.

To set the stage, quantitative responses indicated that participants were more likely than a national sample to be able to identify AI in common applications, suggesting that they had a sufficient understanding of AI to provide ratings of the appropriateness, acceptability, and feasibility of such a tool prior to CORE-MI being introduced. As such, therapists and clinical leadership initially reported moderately high perceptions of the acceptability, appropriateness, and feasibility of an automated fidelity tool for supervision. Participants also reported that they regularly used the types of devices needed for an automated tool in their work,. After discussion of CORE-MI and proposed adaptations for CBT fidelity feedback, therapists' scores of appropriateness were significantly higher, and clinical leaderships' ratings of acceptability, appropriateness, and feasibility were all significantly more positive. These findings suggest that these providers began as positively predisposed toward the idea of automated fidelity tools, and that review of CORE-MI and proposed adaptations, may have elicited favorable responses above and beyond their initial receptiveness.

Initial focus group discussion also pointed to a need for tools to improve supervision. Supervision was described as neither systematic nor targeted at improving specific skills.., and many participants noted the impact of time constraints on giving and receiving supervision. Given the importance of accurate, consistent, actionable feedback for maintaining and improving skills (Creed et al., 2021; Newman & Kaplan, 2016; Schwalbe, et al., 2014; Tracey et al., 2014), access to more systematic and targeted feedback such as that provided by an automatic fidelity tool may present a benefit.

After CORE-MI and the proposed CBT adaptations were introduced, and consistent with previous studies of supervision tools with automated fidelity ratings (Hirsch et al., 2018; Kuo et al., in review), therapists noted that this type of tool would facilitate professional growth, self-reflection, and core skill development. A subset of participants noted that low scores could cause a therapist to doubt their abilities, but others highlighted that improving scores could foster confidence. The immediacy of the feedback was noted as particularly appealing, especially during the learning or training process with new clinical skills, which is consistent with the broader literature about clinical skill development (Newman & Kaplan, 2016; Schwalbe, et al., 2014; Tracey et al., 2014). Other capabilities were highlighted as particularly useful for improving the specificity and effectiveness of supervision, including the searchable session transcript and the opportunity for asynchronous supervision through time stamped comments. In particular, therapists and their clinical leadership noted that the ability to tag specific moments with comments or questions would allow supervision time to focus on high-priority issues and facilitate communication even outside of scheduled supervision time.

Participants were less certain about AI's ability to rate what they perceived as more nuanced skills including rapport, non-verbal communication, and differences related to cultural diversity. Although previous research indicates that sessions can be recorded without negatively impacting rapport (Briggie et al., 2016; Brown, et al., 2013), participant questions were more focused on whether the AI could identify the subtle cues that signify rapport in session. While machine-learning algorithms rely on audio content rather than video to identify patterns like rapport, participants noted that the video would provide additional context for supervision discussions. Similarly, video review may provide important information about culturally-specific differences including clients with high emotionality or specific turns of phrase. Automated fidelity ratings, like any other clinical tool, are best used when integrated with multiple sources of information. A platform like CORE-MI provides an additional information stream but would not—and should not—replace human judgment.

Several participants raised questions about whether the recordings would present a threat to either client or therapist privacy. These segments of the focus group discussion were in contrast to the survey data, which indicated that most participants were familiar and comfortable with recording sessions. The concerns were most closely related to people outside of the agency (e.g., clients, family members of

clients, hackers) being able to access or share recordings and scores because they would be stored on cloud servers. These concerns highlight important feedback from these stakeholder groups about their comfort with, and understanding of, cloud-based technology; however, CORE-MI and any subsequent adaptations rely on HIPAA-compliant (or in Europe, GDPR-compliant) servers, and all data are encrypted as they are uploaded and downloaded, ensuring confidentiality is maintained. Although it is impossible to know what contributed to the change in participants' ratings of acceptability, appropriateness, and feasibility after focus group discussion, one possibility is that discussion of the concerns that were raised may have corrected misperceptions or reassured participants about specific issues.

Finally, in keeping with user-centered design principles (Norman, 2002), participants identified several desired features that could be included in future iterations of CORE-MI or a CBT adaptation. Participants noted the appeal of CBT-specific feedback, and transparency about how scores were generated. Participants noted that information that provided context for scores (e.g., which parts of the session contributed to a specific score, ways to improve their scores in future sessions) would be particularly valuable for improving their skills. Participants also hoped that future iterations could include identification and labeling of specific CBT interventions used in session, perhaps to make progress notes easier to generate.

In concert with the development of automated tools that advance scalable measurement of EBP fidelity (Gibson et al., 2019; Hirsch et al., 2018; Imel et al., 2019; Kuo et al., under review), these findings represent important stakeholder feedback that may help shape such tools to facilitate uptake (Norman, 2002; Proctor et al., 2009; Weiner et al., 2017) and suggest that with such input, automated fidelity measurement could augment standard supervision practices to better support EBP implementation. As supervision becomes more efficient, scalable, and targeted, the skills of both new and seasoned clinicians have the potential to improve (Anker et al., 2009; Lambert et al., 2001)—and by extension, we may potentiate EBPs in community mental health systems. As mental health systems continue to invest in implementing EBPs, these scalable tools will be necessary to ensure that those treatments are delivered as intended, understand reasons for differences in implementation and treatment outcomes, and ensure that consumers of care have access to treatments that work. In addition, findings broaden our understanding of routine supervision, the tools which are already available, and specific areas in which tools may be used to improve supervision efficiency, clinical skills, and treatment delivery.

This study had methodological limitations that should be noted, and additional research is necessary to extend these findings. The study sample was neither large nor random, so the degree to which statistical generalizations may be drawn is limited. Instead, a small and purposive sample was selected to facilitate analytic generalizations (Leech & Onwueguzie, 2010). Replication or extension of these research findings in other community mental health care contexts and systems would increase confidence in their generalizability, including examination of stakeholder feedback after using such a tool in actual clinical practice. In particular, although differences in qualitative themes were not identified between clinicians and clinical leadership, future research should examine whether these groups differ in their priorities or perceptions of the use of an AI-supported supervision tool, given that supervisors and supervisees may interact different with such a tool. In addition, participants were all employed by organizations in a large public mental health system that has championed EBPs and had participated in in CBT training that required ongoing session recordings for (human rated) fidelity assessment (Creed et al., 2021, 2016a, 2016b). While this may limit generalizability of findings related to openness to recording sessions and feedback, this may also present a pathway for normalizing such practices. Future research should also examine stakeholder feedback from among groups who have not had previous experience with routine recording of sessions; without that further study, caution must be used in generalizing these findings beyond those who have normalized session recording or use of EBPs. Given the improvements in attitudes toward the use of a fidelity and supervision platform after being exposed to a specific example, future research may also examine whether this type of exploration may provide a strategy for engaging stakeholders around innovations to facilitate adoption. Additional research is also necessary to better understand the extent to which AI and machine-learning algorithms are able to capture cultural processes and variations in linguistics related to accent or dialect. Finally, given the paucity of access to targeted supervision and skill evaluation in EBPs, the opportunity for asynchronous, remote supervision may offer a strategy for improving access to high quality EBPs; future research should evaluate whether use of AI-based competence evaluation and supervision tools lead to scalable improvements in clinician skill, service outcomes (e.g., treatment retention), or client outcomes (e.g., decreased symptoms, improved quality of life).

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10488-021-01167-x.

**Author contributions** All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by TAC, DR, RO, SO'C, MT, and TH. The draft of the manuscript was written by TAC, PBK, and ZEI, and all authors commented on versions of the manuscript. All authors read and approved the final manuscript.

**Funding** This research was funded by a grant from the National Institute of Mental Health (R56 MH118550 01).

This report follows the guidelines for conducting and reporting mixed-methods research as described in Leech, N.L. & Onwuegbuzie, A.J. (2010). Guidelines for conducting and reporting mixed research in the field of counseling and beyond. *Journal of Counseling and Development, 88(1),* 61–69.

## Declarations

**Disclosure** Drs. Imel, Atkins, Hirsch, and Narayanan are co-founders and equity stakeholders, and Dr. Creed is an equity stakeholder, in Lyssn.io, Inc., a start-up focused on developing technology to support training, supervision, and quality assurance of evidence-based counseling. Dr. Narayanan is also co-founder with equity stake of Behavioral Signals, a start-up focused on creating technologies for emotional and behavioral machine intelligence.

## References


Anker, M. G., Duncan, B. L., & Sparks, J. A. (2009). Using client feedback to improve couple therapy outcomes: A randomized clinical trial in a naturalistic setting. *Journal of Consulting and Clinical Psychology, 77*(4), 693–704. https://doi.org/10.1037/a0016062

Association of State and Provincial Psychology Boards (ASPPB). (2018). Code of conduct. https://cdn.ymaws.com/www.asppb.net/resource/resmgr/guidelines/code_of_conduct_2020_.pdf

Briggie, A. M., Hilsenroth, M. J., Conway, F., Muran, J. C., & Jackson, J. M. (2016). Patient comfort with audio or video recording of their psychotherapy sessions: Relation to symptomatology, treatment refusal, duration, and outcome. *Professional Psychology: Research and Practice, 47*(1), 66–76. https://doi.org/10.1037/a0040063

Brown, E., Moller, N., & Ramsey-Wade, C. (2013). Recording therapy sessions: What do clients and therapists really think? *Counselling & Psychotherapy Research, 13*(4), 254–262. https://doi.org/10.1080/14733145.2013.768286

Center for Behavioral Health Statistics and Quality. (2018). 2017 National Survey on Drug Use and Health: Detailed tables. Substance Abuse and Mental Health Services Administration, Rockville, MD. https://www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUHDetailedTabs2017/NSDUHDetailedTabs2017.pdf

Charmaz, K. (2014). *Constructing grounded theory*. Sage.

Clark D. M. (2011). Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: the IAPT experience. *International Review of Psychiatry*, *23*(4), 318–327. https://doi.org/10.3109/09540261.2011.606803

Creed, T. A., Crane, M. E., Calloway, A., Olino, T. M., Kendall, P. C., & Wiltsey Stirman, S. (2021). Changes in community clinicians' attitudes and competence following a transdiagnostic cognitive behavioral therapy training. *Implementation Research and Practice*. https://doi.org/10.1177/26334895211030220

Creed, T. A., Frankel, S. A., German, R. E., Green, K. L., Jager-Hyman, S., Taylor, K. P., Adler, A. D., Wolk, C. B., Stirman, S. W., Waltman, S. H., & Williston, M. A. (2016a). Implementation of transdiagnostic cognitive therapy in community behavioral health: The beck community initiative. *Journal of Consulting and Clinical Psychology, 84*(12), 1116. https://doi.org/10.1037/ccp0000105

Creed, T. A., Wiltsey-Stirman, S., Evans, A. C., & Beck, A. T. (2014). A model for implementation of cognitive therapy in community mental health: The beck initiative. *The Behavior Therapist, 37*, 56–64.

Creed, T. A., Wolk, C. B., Feinberg, B., Evans, A. C., & Beck, A. T. (2016b). Beyond the label: Relationship between community therapists' self-report of a cognitive behavioral therapy orientation and observed skills. *Administration and Policy in Mental Health Services Research, 43*(1), 36–43. https://doi.org/10.1007/s10488-014-0618-5

Dedoose Version 7.0.23, web application for managing, analyzing, and presenting qualitative and mixed method research data (2020). SocioCultural Research Consultants, LLC. www.dedoose.com

England, M.J., & Butler A.S. (2015). Committee on developing evidence-based standards for psychosocial interventions for mental disorders. *Nationalacademies.org*

Flemotomos, N., Martinez, V., Gibson, J., Atkins, D., Creed, T. A., & Narayanan, S. (2018). Language features for automated evaluation of cognitive behavior psychotherapy sessions. *Proceedings of Interspeech*, *2018*, 1908–1912. https://doi.org/10.21437/Interspeech.2018-1518

Gibson, J., Atkins, D., Creed, T. A., Imel, Z., Georgiou, P., & Narayanan, S. (2019). Multi-label multi-task deep learning for behavioral coding. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2019.2952113

Glaser, B. (1978). *Theoretical sensitivity: Advances in the methodology of grounded theory*. Sociology Press

Goldberg, S. B., Baldwin, S. A., Merced, K., Caperton, D. D., Atkins, D. C., & Creed, T. A. (2020). The structure of competence: Evaluating the factor structure of the Cognitive Therapy Rating Scale. *Behavior Therapy, 51*, 113–122. https://doi.org/10.1016/J.Beth.2019.05.008

Goldberg, S. B., Rousmaniere, T., Miller, S. D., Whipple, J., Nielsen, S. L., Hoyt, W. T., & Wampold, B. E. (2016). Do psychotherapists improve with time and experience? A longitudinal analysis of outcomes in a clinical setting. *Journal of Counseling Psychology, 63*(1), 1–11. https://doi.org/10.1037/cou0000131

Groenewald, T. (2004). A phenomenological research design illustrated. *International Journal of Qualitative Methods, 3*(1), 42–55. https://doi.org/10.1177/160940690400300104

Hirsch, T., Soma, C., Merced, K., Kuo, P., Dembe, A., Caperton, D. D., Atkins, D. C., & Imel, Z. E. (2018). "It's hard to argue with a computer:" Investigating psychotherapists' attitudes towards automated evaluation. *Designing Interactive Systems*. https://doi.org/10.1145/3196709.3196776

Hollon, S. D., DeRubeis, R. J., Andrews, P. W., & Anderson Thomson, J. (2020). Cognitive therapy in the treatment and prevention of depression: A fifty-year retrospective with an evolutionary coda. *Cognitive Therapy and Research*. https://doi.org/10.1007/s10608-020-10132-1

Imel, Z. E., Pace, B. T., Soma, C. S., Tanana, M., Hirsch, T., Gibson, J., Georgiou, P., Narayanan, S., & Atkins, D. C. (2019). Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. *Psychotherapy, 56*(2), 318–328. https://doi.org/10.1037/pst0000221

Jull, J., Giles, A., & Graham, I. D. (2017). Community-based participatory research and integrated knowledge translation: Advancing the co-creation of knowledge. *Implementation Science, 12*, 150. https://doi.org/10.1186/s13012-017-0696-3

Karlin, B. E., Ruzek, J. I., Chard, K. M., Eftekhari, A., Monson, C. M., Hembree, E. A., Resick, P. A., & Foa, E. B. (2012). Dissemination of evidence-based psychological treatments for posttraumatic stress disorder in the Veterans Health Administration. *Journal of Traumatic Stress*, *23*(6), 663–673. https://doi.org/10.1002/jts.20588

Kuo, P.B., Soma, C. S., Axford, K.E., Hirsch, T., Van Epps, J., & Imel, Z.E. (under review). Do as I say, not as I do: Therapist evaluation of a practice and supervision aid

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology, 69*(2), 159–172. https://doi.org/10.1037/0022-006X.69.2.159

Leech, N. L., & Onwuegbuzie, A. J. (2010). Guidelines for conducting and reporting mixed research in the field of counseling and beyond. *Journal of Counseling and Development, 88*(1), 61–70. https://doi.org/10.1002/j.1556-6678.2010.tb00151.x

Mehr, K. E., Ladany, N., & Caskie, G. I. L. (2010). Trainee nondisclosure in supervision: What are they not telling you? *Counselling and Psychotherapy Research, 10*(2), 103–113. https://doi.org/10.1080/14733141003712301

McHugh, R. K., & Barlow, D. H. (2010). The dissemination and implementation of evidence-based psychological treatments: A review of current efforts. *The American Psychologist, 65*(2), 73–84. https://doi.org/10.1037/a0018121

McLean, R., & Tucker, J. (2013). Evaluation of CIHR's Knowledge Translation Funding Program. http://www.cihrirsc.gc.ca/e/47332.html

Muse, K., & McManus, F. (2013). A systematic review of methods for assessing competence in cognitive-behavioural therapy. *Clinical Psychology Review, 33*(3), 484–499. https://doi.org/10.1016/j.cpr.2013.01.010

Newman, C. F., & Kaplan, D. A. (2016). Supervision essentials for cognitive-behavioral therapy. *American Psychological Association*. https://doi.org/10.1037/14950-000

Norman, D. A. (2002). *The design of everyday things*. Basic Books, Inc.

Olfson, M., & Marcus, S. C. (2010). National trends in outpatient psychotherapy. *American Journal of Psychiatry, 167*, 1456–1463. https://doi.org/10.1176/appi.ajp.2010.10040570

Onwuegbuzie, A. J., & Leech, N. L. (2004). Enhancing the interpretation of significant findings: The role of mixed methods research. *The Qualitative Report*, *9*(4), 770–792. https://doi.org/10.46743/2160-3715/2004.1913

Powell, B. J., Stanick, C. F., Halko, H. M., Dorsey, C. N., Weiner, B. J., Barwick, M. A., Damschroder, L. J., Wensing, M., Wolfenden, L., & Lewis, C. C. (2017). Toward criteria for pragmatic measurement in implementation research and practice: A stakeholder-driven approach using concept mapping. *Implementation Science, 12*(1), 118. https://doi.org/10.1186/s13012-017-0649-x

Proctor, E. K., Landsverk, J., Aarons, G., Chambers, D., Glisson, C., & Mittman, B. (2009). Implementation research in mental health services: An emerging science with conceptual, methodological, and training challenges. *Administration and Policy in Mental Health Services Research, 36*(1), 24–34. https://doi.org/10.1007/s10488-008-0197-4

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., Griffey, R., & Hensley, M. (2011). Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health Services Research, 38*(2), 65–76. https://doi.org/10.1007/s10488-010-0319-7

Ronnestad, M. H., & Skovholt, T. M. (1993). Supervision of beginning and advanced graduate students of counseling and psychotherapy. *Journal of Counseling & Development, 71*(4), 396–405. https://doi.org/10.1002/j.1556-6676.1993.tb02655.x

Schwalbe, C. S., Oh, H. Y., & Zweben, A. (2014). Sustaining motivational interviewing: A meta-analysis of training studies. *Addiction, 109*(8), 1287–1294. https://doi.org/10.1111/add.12558

Tracey, T. G. G., Wampold, B. E., Lichtenberg, J. W., & Goodyear, R. K. (2014). Expertise in psychotherapy: An elusive goal? *American Psychologist, 69*, 218–229. https://doi.org/10.1037/a0035099

Waltman, S. H., Frankel, S. A., & Williston, M. A. (2016). Improving clinician self-awareness and increasing accurate representation of clinical competencies. *Practice Innovations, 1*(3), 178–188. https://doi.org/10.1037/pri0000026

Weiner, B. J., Lewis, C. C., Stanick, C., Powell, B. J., Dorsey, C. N., Clary, A. S., Boynton, M. H., & Halko, H. (2017). Psychometric assessment of three newly developed implementation outcome measures. *Implementation Science, 12*(1), 108. https://doi.org/10.1186/s13012-017-0635-3

Xiao, B., Imel, Z. E., Georgiou, P. G., Atkins, D. C., & Narayanan, S. S. (2015). "Rate My Therapist:" Automated detection of empathy in drug and alcohol counseling via speech and language processing. *PLoS ONE, 10*(12), e0143055. https://doi.org/10.1371/journal.pone.0143055

Young, J., & Beck, A.T. (1980). *Cognitive Therapy Rating Scale: Rating manual*. Unpublished manuscript, Center for Cognitive Therapy, University of Pennsylvania, Philadelphia, PA. https://cdn.ymaws.com/www.academyofct.org/resource/collection/24743CF7-351E-4335-9E93-83F26EF675A3/CTRS_Manual.pdf

Zhang, B., & Dafoe, A. (2019). Artificial Intelligence: American attitudes and trends. Center for the Governance of AI, Future of Humanity Institute, University of Oxford. https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/apptopline.html#considersai