Published in final edited form as:
Behav Ther. 2022 November ; 53(6): 1191–1204. doi:10.1016/j.beth.2022.06.001.

A randomized trial to identify accurate measurement methods for adherence to cognitive-behavioral therapy

Emily M. Becker-Haimes[1,2], **Steven C. Marcus**[1], **Melanie R. Klein**[1], **Sonja K. Schoenwald**[3], **Perrin B. Fugo**[1], **Bryce D. McLeod**[4], **Shannon Dorsey**[5], **Nathaniel J. Williams**[6], **David S. Mandell**[1], **Rinad S. Beidas**[1,7,8,9,10]

[1]Department of Psychiatry, University of Pennsylvania

[2]Hall Mercer Community Mental Health, University of Pennsylvania Health System

[3]Oregon Social Learning Center, Eugene, Oregon, United States

[4]Department of Psychology, Virginia Commonwealth University

[5]Department of Psychology, University of Washington

[6]School of Social Work, Boise State University

[7]Leonard Davis Institute of Health Economics, University of Pennsylvania

[8]Department of Medical Ethics and Health Policy, University of Pennsylvania

[9]Department of Medicine, University of Pennsylvania

[10]Penn Implementation Science Center at the Leonard Davis Institute (PISCE@LDI), University of Pennsylvania

Abstract

Clinician fidelity to cognitive behavioral therapy (CBT) is an important mechanism by which desired clinical outcomes are achieved and is an indicator of care quality. Despite its importance, there are few fidelity measurement methods that are efficient and have demonstrated reliability and validity. Using a randomized trial design, we compared three methods of assessing CBT adherence – a core component of fidelity – to direct observation, the gold standard. Clinicians recruited from 27 community mental health agencies (n = 126; *M* age = 37.69 years, *SD* = 12.84; 75.7% female) were randomized 1:1:1 to one of three fidelity conditions: self-report (n = 41), chart-stimulated recall (semi-structured interviews with the chart available; n = 42), or behavioral rehearsal (simulated role plays; n = 43). All participating clinicians completed fidelity assessments for up to three sessions with three different clients that were recruited from clinicians' caseloads (n = 288; *M* age = 13.39 years *SD* = 3.89; 41.7% female); sessions were also audio-recorded and coded for comparison to determine the most accurate method. All fidelity measures had parallel scales that yielded an adherence maximum score (i.e., the

Correspondence should be addressed to Emily M. Becker-Haimes, Department of Psychiatry, University of Pennsylvania, Perelman School of Medicine, 3535 Market Street, 3rd floor, Philadelphia, PA 19104, United States. emily.haimes@pennmedicine.upenn.edu.

Trial Registration: ClinicalTrials.gov identifier: NCT02820623

Conflicts of Interest: Dr. Beidas discloses that she receives royalties from Oxford University Press and has consulted for Camden Coalition of Healthcare Providers. All other authors have no conflict of interest to disclose.

highest-rated intervention in a session), a mean of techniques observed, and a count total of observed techniques. Results of three-level mixed effects regression models indicated that behavioral rehearsal produced comparable scores to observation for all adherence scores (all *p*s > .01), indicating no difference between behavioral rehearsal and observation. Self-report and chart-stimulated recall overestimated adherence compared to observation (*p*s <.01). Overall, findings suggested that behavioral rehearsal indexed CBT adherence comparably to direct observation, the gold-standard, in pediatric populations. Behavioral rehearsal may at times be able to replace the need for resource-intensive direct observation in implementation research and practice.

Keywords

youth mental health; cognitive-behavioral therapy; fidelity; adherence

Introduction

Cognitive-behavioral therapy (CBT) is the leading psychosocial intervention for many youth psychiatric disorders based on its effectiveness and cost (Dorsey et al., 2017; Higa-McMillan et al., 2016; Hofmann et al., 2012; Weisz et al., 2017). Many systems have invested in CBT implementation over the past decade to improve the quality of mental healthcare, and there are increasing efforts to empirically examine the effectiveness of these system level investments (Beidas et al., 2019; Hoagwood et al., 2014; Regan et al., 2017). Measuring fidelity to CBT delivery is a barrier for both quality assurance and implementation studies. Clinician *fidelity* refers to whether CBT is delivered as intended, with clinician *adherence* to CBT techniques theorized to be the core fidelity component by which desired clinical outcomes are achieved (Chiapa et al., 2015; Hogue et al., 2008; Huey Jr et al., 2000; McLeod et al., 2013; Miller & Rollnick, 2014; Proctor et al., 2011; Schoenwald et al., 2011). We define adherence to CBT as the degree to which a clinician employs CBT techniques with the breadth and depth intended (Perepletchikova et al. 2007). While key questions remain about how and when clinician CBT adherence produces improved client outcomes (e.g., Southam-Gerow et al., 2021; Rapley & Loades, 2018), clinician CBT adherence is a primary target for implementation efforts and quality improvement efforts in mental healthcare (McLeod et al., 2013; Proctor et al., 2011).

To bring evidence-based treatments such as CBT to scale to optimize youth mental health care across diverse settings, the field needs pragmatic methods (Glasgow & Riley, 2013; Lewis et al., 2015) to measure clinician adherence in practice and in implementation and effectiveness trials. Currently, the gold-standard approach for measuring clinician CBT adherence relies on observational methods (Rodriguez-Quintana & Lewis, 2018). Such methods most commonly take the form of independent coding of clinician behavior during live observation or review of therapy session recordings. Studies utilizing direct observation to measure clinician CBT adherence suggest that it can be implemented reliably (McLeod et al., 2015), can identify differences in CBT adherence across community and research contexts (e.g., McLeod et al., 2017; Smith et al., 2017), and can support efforts to examine possible links between adherence and clinical outcomes (e.g., Southam-Gerow et al., 2021). However, direct observational methods are expensive (Simons et al., 2013) and require expertise and infrastructure that most community settings do not have (e.g., recording equipment or live observation windows, availability of expert supervisors to observe sessions, quantify adherence, and provide feedback to clinicians). It is also burdensome for researchers, clinicians, and clients in implementation trials (e.g., clients must agree to be recorded and observed in therapy, and time from therapy sessions sometimes may be taken to undergo consent procedures).

This randomized trial compared three methods – self-report, chart-stimulated recall, and behavioral rehearsal – for assessing clinician CBT adherence in youth mental health practice. Each was compared with direct observation to identify the most accurate method for indexing CBT adherence in pediatric mental health care (see Beidas et al., 2016). While self-reported adherence is inexpensive and relatively easy to collect (Schoenwald et al., 2011), a robust literature details the limitations of clinicians' ability to accurately self-rate their adherence (e.g., Brosan et al. 2008; Creed et al., 2014; Hurlburt et al., 2010; Martino et al., 2009). As such, we enhanced the self-report condition by using a measure that was developed in collaboration with community clinicians (Becker-Haimes et al., 2021) and included training for clinicians in how to self-rate accurately prior to its use. Chart-stimulated recall is an interviewing technique used to measure the content and processes that occur in a clinical encounter. Chart-stimulated recall is used in other areas of medicine and consists of a brief structured interview between a trained research assistant or supervisor and the clinician about their clinical practice; clinicians use the client chart to prompt their memory of their treatment delivery. Behavioral rehearsal, also known as the standardized patient method, is a well-established method in medicine for evaluating physician behavior via role play (Epstein, 2007) that has also been used successfully in mental health settings (e.g., Beidas et al., 2014). Traditional behavioral rehearsal, however, does not assess adherence with *specific* clients. We created a novel version of behavioral rehearsal in which the clinician demonstrates their use of CBT in a specific client session with a trained interviewer who rates CBT adherence.

Both chart-stimulated recall and behavioral rehearsal are more resource-intensive than self-report but may yield more accurate measures of adherence with less burden than direct observation. We hypothesized that chart-stimulated recall and behavioral rehearsal would more accurately measure clinician adherence than self-report, with chart-stimulated recall yielding better accuracy than both self-report and behavioral rehearsal.

Material and Methods

The City of Philadelphia (Approval #2016–24) and the University of Pennsylvania (Approval #834079) Institutional Review Boards approved this study. Informed written or verbal consent or assent was obtained from all clinicians, guardians, and youth.

Procedure

Clinician recruitment occurred via staff meetings at 27 publicly funded mental health clinics whose leadership provided approval for procedures. All agencies were from the Philadelphia, PA tristate area; most were members of a publicly funded mental health system that has supported the widespread implementation of CBT for youth (Beidas et al., 2013; Beidas et al., 2019; Powell et al., 2016). We advertised the study to 53 agencies; 40 expressed interest, 39 were eligible (i.e., had clinicians delivering CBT to study eligible youth), and 27 participated (i.e., allowed clinician recruitment to occur from their site).

Study staff described the study to clinicians; interested clinicians were consented and randomized 1:1:1 to each experimental condition. Eligible clinicians were those who could identify at least 3 eligible client sessions within the next month. Eligible sessions included those in which the clinician intended to use at least 1 of 12 CBT techniques, the client was aged 7–24 (in Philadelphia, pediatric services may extend to age 24), the session occurred in English, and the client had a legal guardian who could provide consent for client participation (e.g., clients in foster care were ineligible; see Beidas et al., 2016 for detailed eligibility criteria). Clinicians also had to plan to direct their intervention strategies to the youth client for at least ten minutes of the session. As such, youth under the age of 7 were excluded from study participation, as leading evidence-based treatment models for such youth are often directed primarily toward caregivers (e.g., Thomas et al., 2017). We excluded first-session encounters, as these often focus more on assessment and rapport building than CBT delivery.

Following randomization, study staff worked with clinicians to identify and record sessions (*M* client per clinician = 2.29). Clinicians were compensated $25 for their time providing their schedule. Clinicians then approached clients and guardians directly, provided a study overview, and assessed interest. Study staff reviewed all elements of informed consent (and assent for youth under 18) with clients who expressed initial interest and managed the audio recording process for clients who consented to participate. Clients were compensated $10. All sessions were audio-recorded and uploaded to REDCap (Harris et al., 2009), a HIPAA compliant platform.

Clinicians randomized to self-report were asked to complete the self-report measure within 48 hours of the clinical encounter. For the behavioral rehearsal and chart-stimulated recall conditions, clinicians in completed measurements within 1 week of recorded encounters, with few exceptions (e.g., holidays; *M* = 5.76 days, *SD* = 5.45). Study staff who coded session recordings were masked to content shared by clinicians in each condition. Clinicians were compensated $50 per hour for research activity participation (*M* = $87.77, *SD* = 15.48).

Participants

Clinicians.—Clinicians (*n* = 126) were recruited from 27 community mental health agencies in the Philadelphia tristate region between October 2016 and May 2020. Of enrolled clinicians, 103 enrolled at least one client included in analysis (see Figure 1 for CONSORT Flow diagram); these clinicians were comparable to those who did not enroll a client/s on background characteristics. Included clinicians (*M* age = 37.69 years, *SD* = 12.84) 78 (75.7%) identified as female, 98 (95.1%) held master’s degrees, and 73 (70.9%) as White, 17 (16.5%) as Black or African American, 5 (4.9%) as Asian or Pacific Islander, 2 (1.9%) as more than one race, and 6 (5.8%) did not disclose; 6 (5.8%) identified as Hispanic or Latinx. Table 1 shows clinician demographics.

Clients.—Clients (*n* = 304) were aged 7–24 receiving therapy from clinicians enrolled in the trial. Sixteen clients were excluded from analysis (e.g., session was not in English, client withdrew; see Figure 1). Clients were demographically and diagnostically similar to those typically seen in community settings (Beidas et al., 2019; Substance Abuse and Mental Health Services Administration, 2019). Included clients (*n* = 288) averaged 13.39 years old (*SD* = 3.89). 120 (41.7%) identified as female and 166 (57.6%) male, and 121 (42%) identified as Black or African American, 94 (32.6%) as White, 5 (1.7%) as Asian, 1 (0.4%) as Native American or Alaska Native, 1 (0.4%) as Native Hawaiian or other Pacific Islander, 7 (2.4%) as some other race, 25 (8.7%) as more than one race, 34 (11.8%) did not disclose; 64 (22.2%) Hispanic or Latinx. Client diagnoses, as reported by their clinicians, varied; 53.1% had a primary internalizing and 39.9% had a primary externalizing disorder; 58.9% had one or more co-morbid diagnoses. Table 2 shows client demographics.

Measures

Direct Observation.—The Therapy Process Observational Coding System- Revised Strategies (TPOCS-RS) Scale (McLeod et al., 2015) was the gold-standard observational coding system used to capture clinician CBT adherence on a range of youth CBT techniques for all audio-recorded sessions. The TPOCS-RS has been used in multiple studies and shows good internal consistency and validity (McLeod et al., 2015; McLeod & Weisz, 2010; Smith et al., 2017). The TPOCS-RS yields individual extensiveness scores for 12 CBT techniques (e.g., cognitive restructuring, relaxation) on a 7-point Likert scale (1 = Not Present to 7 = Extensively). Extensiveness is defined as a combination of the frequency and thoroughness of intervention delivery. Detailed descriptions of the 12 CBT techniques of interest in this study are available in Supplemental File 1.

The TPOCS-RS yields several aggregate CBT adherence scores, each of which captures slightly different ways of indexing adherence (Beidas et al., 2017; McLeod et al., 2015): (1) a Maximum CBT score, defined as the highest coded intervention technique across all 12 possible interventions in a given session (possible range = 1 [*not present*]-7 [*extensively present*]), (2) a Mean CBT score, defined as the average of all coded interventions in a given session (possible range = 1 [*not present*]-7 [*extensively present*]), and (3) a CBT Count Total of Techniques, defined as the total number of discrete CBT intervention techniques coded as present in the session; (possible range = 0 [*no CBT interventions present*]-12 [*all possible interventions present*]). Both depth and breadth of CBT delivery are important to capture in an adherence index (Garland et al., 2006). These three indices vary in whether they primarily capture depth of CBT delivery (the Maximum CBT Score), capture a combination of depth and breadth (the Mean CBT Score), or primarily capture breadth of delivery (CBT Count Total). It is important to note for this last outcome that higher CBT Count Total scores are not necessarily indicative of better or more adherent CBT practice; for example, a clinician delivering a single intervention with high extensiveness would receive a CBT Count Total Score of 1, whereas a clinician who delivered 4 interventions with low extensiveness would receive a CBT Count Total Score of 4.

All 288 sessions were coded by 1 of 11 raters. Raters comprised a mix of clinical research coordinator staff, clinical psychology graduate students, and doctoral level raters. Before coding, all raters independently coded at least 15 certification sessions and achieved established interrater reliability benchmarks (item-level intraclass correlation coefficients [ICCs (2,2)] > .60) against gold-standard trainer's ratings from the TPOCS-RS measure, using established procedures (McLeod et al., 2015). Raters attended biweekly meetings to prevent drift led by an expert coder, in which the codebook was regularly reviewed, and any coding questions were raised and discussed as a full team. Several times per year, the full team also collaboratively coded a single session to ensure calibration. Forty-nine percent of sessions were double coded by a doctoral level expert coder. Interrater agreement on the TPOCS-RS was high on all techniques (item ICCs ranged from .76-.95). Average time to code each session was 47.32 minutes (*SD* = 16.56), slightly longer than recorded session lengths (*M* = 43.24; *SD* = 11.16).

Supplemental Files 1–3 include copies of the self-report, chart-stimulated recall, and behavioral rehearsal instruments.

Self-Report.—Clinicians completed the TPOCS Self-Reported Therapist Intervention Fidelity for Youth (TPOCS-SeRTIFY; Becker-Haimes et al., 2021; Beidas et al., 2016). The TPOCS-SeRTIFY provides operational definitions for each CBT technique and has a companion 30-minute training and rating manual that includes sample vignettes of clinician behaviors and how those vignettes should be rated. The TPOCS-SeRTIFY asks clinicians to self-rate their use of the same techniques as the TPOCS-RS using parallel Likert scales. Mean time for clinicians to complete the TPOCS-SeRTIFY was 8.74 minute per session (*SD* = 6.19). Clinicians reported that TPOCS-SeRTIFY responses accurately reflected what they did in session as 5.02 (*SD* = 1.08, Range = 2–7) on a scale of 1 (Not confident at all) to 7 (Very confident). Initial psychometric analysis of this measure suggested strong item performance and preliminary construct validity, with strong concordance with another established self-report measure (Becker-Haimes et al., 2021).

Chart-Stimulated Recall.—Standard chart-stimulated recall methodology (Guerra et al., 2007) was adapted to assess the 12 techniques measured on the TPOCS-RS using a parallel scale. Research staff who administered chart-stimulated recalls (a mix of clinical research coordinator staff, clinical psychology graduate students, and postdoctoral fellows) were first trained in the TPOCS-RS, met reliability on 4 audio-recorded chart-stimulated recalls, and completed a mock chart-stimulated recall and 2 supervised ones in the field before independent administration. Research staff attended monthly supervision to prevent drift. 23% of sessions were double coded for reliability by a postdoctoral fellow with CBT expertise; interrater agreement was excellent (ICCs for individual techniques ranged .90-.98). Chart-stimulated recalls were administered at the clinicians' agency. Client information shared with the interviewer was deidentified; only the clinician interacted with the client chart. Mean chart-stimulated recall administration time was 19.01 minutes per session (*SD* = 5.01). The mean time to score each chart-stimulated recall was 8.4 minutes; the median time for completing both administration and scoring was 21.3 minutes. Clinicians reported that chart-stimulated recall accurately reflected what they did in session as 5.19 (*SD* = 0.94, Range = 3–7) on a scale of 1 (Very Poor) to 7 (Excellent).

Behavioral Rehearsal.—Traditional behavioral rehearsal refers to a standardized role-play between a clinician and a member of the research team acting as their client (Beidas et al., 2014). However, we adapted this traditional paradigm to align behavioral rehearsal with the session-specific targets of direct observation measurement. Specifically, we instructed clinicians to engage in a semi-structured role play with a trained interviewer for up to 15 minutes and "demonstrate how they used CBT in their session." The interviewer acted as the client and later rated CBT adherence for techniques demonstrated on items parallel to the TPOCS-RS. Clinicians were asked to condense their use of CBT in the role play and were encouraged to role play multiple session interactions (e.g., at different time points throughout the session, with different members of the family system present in session), if needed, to role play all CBT techniques used. The role play was conducted at the clinicians' agency, prior to which, clinicians provided brief, deidentified information to interviewers about their client to help guide the role play practice. Behavioral rehearsal interviewers (a mix of clinical research coordinator staff, clinical psychology graduate students, and postdoctoral fellows) were trained in the TPOCS-RS, met reliability on coding techniques on 2 audio-recorded behavioral rehearsals, and completed a minimum of 2 mock and at least 1 supervised behavioral rehearsal in the field before independent administration. All behavioral rehearsals were audio recorded. Interviewers attended monthly supervision to prevent drift; 42% of audio-recorded behavioral rehearsals were double coded for reliability by a postdoctoral fellow with CBT expertise. Interrater agreement was excellent (ICCs for individual techniques ranged .84–1.00). The mean administration time to set up and conduct the role-play was 17.26 minutes/session (*SD* = 4.17). The mean time to score each behavioral rehearsal was 23.7 minutes; the median time for completing both administration and scoring was 35.2 minutes. Clinicians reported that behavioral rehearsal accurately reflected what they did in session as 4.77 (*SD* = 1.28, Range = 1–7) on a scale of 1 (Very Poor) to 7 (Excellent).

Statistical Analysis

Overall rates of missing data were low (0.02%). No variable was missing more than 1.4% of its values and data were missing completely at random (Little's MCAR X^2 = 172.8, p = .99) (Little, 1988). Preliminary analyses compared baseline demographic and clinical characteristics of clinicians and youth across conditions to check randomization, using t-tests for continuous and X^2 tests for categorical variables. There were no differences in baseline demographic and clinical characteristics of clinicians and youth on any examined variable between conditions (see Tables 1 and 2), indicating successful randomization.

As described above, fidelity outcomes of interest were the three indices of adherence calculated from all three measurement conditions and from the TPOCS-RS direct observation codes: (1) a Maximum CBT score (i.e., the highest coded intervention technique across all 12 possible interventions in a session), (2) a Mean CBT score (i.e., the average of all coded interventions in a session, and (3) a CBT Count Total (i.e., the total number of discrete CBT intervention techniques coded as present in a session). Because the data were nested (Level 1 = clients, Level 2 = clinicians, Level 3 = agency), three-level regressions with random intercepts in SAS Version 9.4 calculated the least squares mean paired difference between scores obtained through direct observation with the TPOCS-RS and the scores from each condition. Primary outcomes were the significance tests of the paired differences between direct observation scores and each condition. A *non-significant p* value indicated that the condition produced comparable scores to those as direct observation. To facilitate interpretation of the magnitude of the differences between each condition and direct observation scores, we calculated Cohen's d using recommendations for calculating effect sizes within the context of mixed models (Feingold, 2013); effect size interpretation followed conventional guidelines, such that Cohen's d values of of .20, .50, and .80 indicated small, medium, and large effects, respectively (Cohen, 1988).

We also examined how each condition performed relative to direct observation as compared to the other two measurement conditions (i.e., self-report vs. chart-stimulated recall, self-report vs. behavioral rehearsal, chart-stimulated recall vs. behavioral rehearsal). We used a Tukey-Kramer multiple-comparison test to examine whether the paired mean difference scores between each condition and direct observation significantly differed by condition (e.g., if the magnitude of the difference between direct observation and self-report differed from the magnitude of the difference between direct observation and chart-stimulated recall) for all adherence scores and for all pair-wise condition comparisons. Given the number of models, we set the alpha level at .01 for all analyses to minimize risk of Type 1 error, in addition to the adjustments described.

Results

Primary Outcomes.

Table 3 shows results of the comparison of each condition's performance relative to direct observation.

Self-Report.—The least squares mean paired difference between self-report and direct observation was significantly different from zero for all scores. Specifically, self-report yielded higher estimates of the maximum observed CBT extensiveness score (Maximum CBT $M = 5.28$, $SD = 1.17$) than those obtained via direct observation ($M = 3.42$, $SD = 1.55$; $M_{diff} = -1.84$, Cohen's d = 1.34, $p < .001$), and yielded higher estimates of the mean of all observed CBT techniques (Mean CBT $M = 3.91$, $SD = 0.82$) compared to direct observation ($M = 2.77$, $SD = 0.72$; $M_{diff} = -1.15$, Cohen's d = 1.49, $p < .0001$). Self-report scores estimated that clinicians used an average of 7.17 discrete CBT interventions per session ($SD = 2.59$), whereas direct observation estimated an average of 3.15 CBT interventions per session ($SD = 1.84$; $M_{diff} = -4.02$, Cohen's d = 1.79, $p < .001$).

Chart-Stimulated Recall.—Contrary to hypotheses, chart-stimulated recall also produced adherence scores higher than those obtained by direct observation for all adherence indices. Specifically, chart-stimulated recall yielded higher estimates of the maximum observed CBT extensiveness score (Maximum CBT $M = 4.53$, $SD = 3.69$) than those obtained via direct observation ($M = 3.69$, $SD = 1.36$; $M_{diff} = -0.83$, Cohen's d = 0.62, $p < .001$), and yielded higher estimates of the mean of all observed CBT techniques (Mean CBT $M = 3.18$, $SD = 0.63$) compared to direct observation ($M = 2.80$, $SD = 0.68$; $M_{diff} = -0.38$, Cohen's d = 0.58, $p = .002$). Chart-stimulated recall estimated that clinicians used an average of 4.38 discrete

CBT interventions per session ($SD = 1.74$), whereas direct observation estimated an average of 3.15 CBT interventions per session ($SD = 1.40$; $M_{diff} = 1.23$, Cohen's d = 0.77, $p <.001$).

Behavioral Rehearsal.—Consistent with hypotheses, behavioral rehearsal produced adherence scores comparable to those obtained by direct observation for all adherence scores. Specifically, behavioral rehearsal yielded non-significantly different estimates of the maximum observed CBT extensiveness score (Maximum CBT $M = 4.09$, $SD = 1.48$) than those obtained via direct observation ($M = 3.72$, $SD = 1.59$; $M_{diff} = -0.34$, Cohen's d = 0.22, $p = .08$), and non-significantly different estimates of the mean of all observed CBT techniques (Mean CBT $M = 3.13$, $SD = 0.93$) compared to direct observation ($M = 2.82$, $SD = 0.70$; $M_{diff} = -0.30$, Cohen's d = 0.36, $p =.02$). Behavioral rehearsal estimated that clinicians used an average of 3.10 discrete CBT interventions per session ($SD = 1.58$), whereas direct observation estimated an average of 3.21 CBT interventions per session ($SD = 1.78$; $M_{diff} = 0.12$, Cohen's d = 0.07, $p = .75$).

Table 4 shows results examining how each condition performed relative to direct observation when compared to the other two measurement conditions.

Self-Report versus Chart-Stimulated Recall.—Comparison of the relative performance of self-report and chart-stimulated recall indicated that, consistent with hypotheses, chart-stimulated recall yielded scores closer to those obtained by direct observation than did self-report on all outcomes. When compared to chart-stimulated recall, self-report yielded scores, on average, 1.01 points higher on the Maximum CBT score ($p < .001$), 0.77 points higher on the Mean CBT score ($p <.0001$) and indicated the presence of 2.79 more discrete CBT techniques per session on the CBT Count Total ($p = .002$).

Self-Report versus Behavioral Rehearsal.—Comparison of the relative performance of self-report and behavioral rehearsal indicated that, consistent with hypotheses, behavioral rehearsal yielded scores closer to those obtained by direct observation than did self-report on all outcomes. When compared to behavioral rehearsal, self-report yielded scores, on average, 1.50 points higher on the Maximum CBT score ($p <.0001$), 0.85 points higher on the Mean CBT score ($p <.0001$) and indicated the presence of 4.14 more discrete CBT techniques per session on the CBT Count Total score ($p = .002$).

Chart-Stimulated Recall versus Behavioral Rehearsal.—Contrary to hypotheses, there was no difference in the relative performance of behavioral rehearsal and chart-stimulated recall on any outcomes, indicating that scores obtained by chart-stimulated recall were not significantly different than those obtained by behavioral rehearsal for the Maximum CBT Score (M_{diff} between chart-stimulated recall and behavioral rehearsal $= 0.49$ $p =.15$), the Mean CBT score (M_{diff} between chart-stimulated recall and behavioral rehearsal $= -0.30$, $p =.90$, and the CBT Count Total of Techniques Present (M_{diff} between chart-stimulated recall and behavioral rehearsal $= 0.12$, $p = .02$).

Discussion

This randomized controlled trial compared the accuracy of three adherence measurement methods to direct observation for youth CBT in community mental health settings. Results have clear implications for research and practice. Notably, our adapted behavioral rehearsal methodology, which consisted of brief, semi-structured role plays (< 15 minutes), provided comparable estimates of one critical component of fidelity - clinician CBT adherence - compared to direct observation. Behavioral rehearsal methodology, when conducted in a semi-structured format, is thus a potentially less resource intensive for organizations and clients but comparable method for assessing clinician CBT adherence compared to direct observation. This has implications for reducing client burden in both quality assurance and implementation trials. In particular, use of semi-structured behavioral rehearsal methodology can potentially reduce or eliminate the need for client therapy sessions to be recorded and coded to reliably monitor implementation outcomes or support quality assurance procedures. It also does not require obtaining client consent, as behavioral rehearsals were conducted feasibly without the use of identifiable client information.

Contrary to hypothesis, findings from the chart-stimulated recall arm, which consisted of brief semi-structured interviews with the clinicians about the intervention techniques used in session with their youth clients and the clinician uses their client’s chart to prompt recall, suggested that the chart-stimulated recall overestimated clinician adherence, with medium effect sizes. One potential explanation for this is that, in practice, chart-stimulated recall relied to a degree on clinicians self-reporting the intervention techniques used to a trained interviewer. This suggests that chart-stimulated recall may, to a degree, be limited by the same self-reporting biases that have historically limited the accuracy of self-reported adherence measures (e.g., recall bias, social desirability; Killeen et al., 2004). However, chart-stimulated recall did meaningfully outperform the self-report condition at indexing adherence, suggesting that the use of a trained interviewer and availability of the chart enhances the accuracy of clinician adherence reporting.

Self-report, in which clinicians rated their intervention techniques used on a structured form following their recorded session, yielded the scores most discrepant from direct observation of all conditions, with effect sizes suggesting a large effect for the magnitude of the difference between self-report and direct observation. Notably, the self-report condition identified approximately 4 more intervention techniques per session than what was coded on the TPOCS-RS in direct observation. This suggests that, not only did clinicians overestimate the extensiveness of their CBT delivery, but clinicians also report using many more intervention techniques than are observed by independent coders. That said, the TPOCS-RS, the direct observation coding system used in this study, is designed to code explicit clinician behaviors observed in session; coders are instructed not to code CBT techniques they believe a clinician is attempting to deliver unless it reaches a certain threshold (i.e., they should not infer what a clinician is doing). This could potentially explain why clinicians in this study reported perceiving themselves as delivering more interventions than what was observed by coders. However, findings converge with literature highlighting the tendency of clinicians to overestimate their adherence (Brosan et al., 2008; Carroll et al., 1998; Hogue et al., 2015; Hurlburt et al., 2010; Martino et al., 2009). This trial extends this work by demonstrating that brief training for clinicians in how to self-rate does not produce self-reported adherence scores comparable to direct observation (SR). An important next step will be to identify whether we can identify predictors (e.g., clinician CBT knowledge) of clinicians who can more accurately self-rate their adherence. In addition, future work should also identify whether there is variability in self-rating accuracy as a function of specific intervention characteristics (e.g., observability, salience, complexity). Such work could inform the development of algorithms to make data from self-report or chart-stimulated recall more useful by adjusting scores obtained by self-report or chart-stimulated recall as a function of these predictors.

Optimizing our ability to feasibly measure clinician adherence also has implications for clinical practice. There is increasing burden on agencies to demonstrate quality of care, particularly with value-based care payments; identifying adherence measurement tools that can be integrated into clinical supervision and quality assurance efforts is an important next step. While only the semi-structured behavioral rehearsal produced estimates of clinician CBT use comparable to direct observation, relative comparisons between conditions did not suggest that behavioral rehearsal meaningfully outperformed chart-stimulated recall. Furthermore, while behavioral rehearsal represented less administration time ($M = 17$ minutes) than recording or observing a full session, additional time was needed for training and scoring for accurate rating of the behavioral rehearsal condition; this was especially true for research team members who did not have prior formal training in CBT (i.e., they tended to take more time to rate behavioral rehearsals than did those with formal CBT training). Overall, the median time taken to administer and score behavioral rehearsals was 35.2 minutes, which was only 12 minutes less than the average time taken to obtain adherence scores via direct observation. Thus, while behavioral rehearsal may offer time reductions when used for internal quality assurance or training purposes (e.g., with an expert supervisor), its overall administration cost to a research team evaluating fidelity may be somewhat higher than that of chart-stimulated recall. In addition, while behavioral rehearsals may reduce burdens on clients (who do not need to be recorded) and organizations (who do not need to cover the costs of technological infrastructure required for recording sessions), the behavioral rehearsal may incur *more* costs to clinician time than direct observation, as the clinician must additionally complete the role play. Clinician costs may be offset by the opportunity to receive feedback from role-play observers to improve clinical practices when employed by supervisors in the context of quality assurance and improvement; however, how and when to optimally utilize something like behavioral rehearsal to support quality improvement remains an area for future research. Formal economic evaluation and stakeholder preference for each of the three conditions are forthcoming and will shed light on the cost-effectiveness and acceptability of each approach to inform potential integration into supervisory and quality assurance structures. Additional work also is underway to examine how self-report, chart-stimulated recall, and behavioral rehearsal perform in their ability to index clinician competence (i.e., how skillfully and responsively clinicians deliver CBT interventions), which is another critical component of clinician fidelity (Perepletchikova et al., 2007).

Results should be interpreted within the context of study limitations. The primary study limitation is the relatively low use of CBT across the sample; it is possible that measurement methods may perform differently in clinical settings where CBT delivery is more commensurate with levels seen in efficacy trials. For example, self-report and chart-stimulated recall may more accurately index adherence among clinicians who use CBT techniques regularly at high levels. Related, we initially planned only to recruit clinicians who had been formally trained in CBT through intensive, city-sponsored training initiatives; however, due to recruitment challenges, we widened the scope to any clinician who self-reported having training in CBT. This may have impacted our findings; examining how clinician CBT training background moderates the performance of each fidelity condition is an important area for future inquiry. In addition, as we excluded sessions that were explicitly targeted towards caregivers, we thus are unable to examine how each measurement condition may have performed in non-youth focused sessions. An additional limitation of note is that the research staff conducted the chart-stimulated recall and behavioral rehearsal fidelity measurement conditions. To optimally understand which method can best support quality assurance efforts, an important area for future research will be to examine how each of these measurement methods performs when measures are administered by agency-based supervisors or administrators. Finally, while we anticipate that the results for each measurement method are likely to generalize to other populations (e.g., CBT use with adults; Young & Beck, 1980) or interventions (e.g., interpersonal psychotherapy for adolescents; Mufson & Sills, 2009); our focus on CBT adherence in a youth population as an exemplar necessitates further work in other populations. In addition, regardless of the effectiveness of novel fidelity methods, direct observation may remain a useful care quality tool in some instances (e.g., home-based therapy).

This study also has notable strengths. This is the first large-scale randomized trial to demonstrate that a less intensive fidelity measure (behavioral rehearsal) can provide comparable estimates of clinician adherence, a key metric of implementation efforts' success, to direct observation which has implications for implementation research and practice.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

We are grateful for the support that the Department of Behavioral Health and Intellectual Disability Services has provided us to conduct this work within their system, for the Evidence Based Practice and Innovation (EPIC) group, and for the partnership provided to us by participating agencies, clinicians, and clients.

Funding:

This work was supported by National Institute of Mental Health Grant R01 MH108551 (PI: Beidas).

Data Availability:

The data that support the findings of this study are available from the corresponding author, EBH, upon reasonable request.

References

Becker-Haimes EM, Klein MR, McLeod BD, Schoenwald SK, Dorsey S, Hogue A, Fugo PB, Phan ML, Hoffacker C, & Beidas RS (2021). The TPOCS-self-reported Therapist Intervention Fidelity for Youth (TPOCS-SeRTIFY): A case study of pragmatic measure development. Implementation Research and Practice, 2, 1–9. 10.1177/2633489521992553

Beidas RS, Aarons GA, Barg F, Evans AC, Hadley T, Hoagwood KE, Marcus SC, Schoenwald SK, Walsh LM, & Mandell DS (2013). Policy to implementation: evidence-based practice in community mental health–study protocol. Implementation Science, 8, 38. 10.1186/1748-5908-8-38 [PubMed: 23522556]

Beidas RS, Becker-Haimes EM, Adams DR, Skriner L, Stewart RE, Wolk CB, Buttenheim AM, Williams NJ, Inacker P, Richey E, & Marcus SC (2017). Feasibility and acceptability of two incentive-based implementation strategies for mental health therapists implementing cognitive-behavioral therapy: a pilot study to inform a randomized controlled trial. Implementation Science, 12(1), 148. 10.1186/s13012-017-0684-7 [PubMed: 29246236]

Beidas RS, Cross W, & Dorsey S. (2014). Show me, don't tell me: behavioral rehearsal as a training and analogue fidelity tool. Cognitive and Behavioral Practice, 21(1), 1–11. 10.1016/j.cbpra.2013.04.002 [PubMed: 25382963]

Beidas RS, Maclean JC, Fishman J, Dorsey S, Schoenwald SK, Mandell DS, Shea JA, McLeod BD, French MT, Hogue A, Adams DR, Lieberman A, Becker-Haimes EM, & Marcus SC (2016). A randomized trial to identify accurate and cost-effective fidelity measurement methods for cognitive-behavioral therapy: project FACTS study protocol. BMC Psychiatry, 16(1), 323. 10.1186/s12888-016-1034-z [PubMed: 27633780]

Beidas RS, Williams NJ, Becker-Haimes E, Aarons G, Barg F, Evans A, Jackson K, Jones D, Hadley T, Hoagwood K, Marcus SC, Neimark G, Rubin R, Schoenwald S, Adams DR, Walsh LM, Zentgraf K, & Mandell DS (2019). A repeated cross-sectional study of clinicians' use of psychotherapy techniques during 5 years of a system-wide effort to implement evidence-based practices in Philadelphia. Implement Science, 14(1). 10.1186/s13012-019-0912-4

Brosan L, Reynolds S, & Moore RG (2008). Self-evaluation of cognitive therapy performance: do therapists know how competent they are? Behavioural and Cognitive Psychotherapy, 36(5), 581. 10.1017/S1352465808004438

Carroll K, Nich C, & Rounsaville B. (1998). Utility of therapist session checklists to monitor delivery of coping skills treatment for cocaine abusers. Psychotherapy Research, 8(3), 307–320. 10.1080/10503309812331332407

Chiapa A, Smith JD, Kim H, Dishion TJ, Shaw DS, & Wilson MN (2015). The trajectory of fidelity in a multiyear trial of the family check-up predicts change in child problem behavior. Journal of Consulting and Clinical Psychology, 83(5), 1006–1011. 10.1037/ccp0000034 [PubMed: 26121303]

Creed TA, Wolk CB, Feinberg B, Evans AC, & Beck AT (2016). Beyond the label: Relationship between community therapists' self-report of a cognitive behavioral therapy orientation and observed skills. Administration and Policy in Mental Health and Mental Health Services Research, 43(1), 36–43. 10.1007/s10488-014-0618-5 [PubMed: 25491201]

Cohen J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum

Dorsey S, McLaughlin KA, Kerns SE, Harrison JP, Lambert HK, Briggs EC, Revillion Cox J, & Amaya-Jackson L. (2017). Evidence base update for psychosocial treatments for children and adolescents exposed to traumatic events. Journal of Clinical Child and Adolescent Psychology, 46(3), 303–330. 10.1080/15374416.2016.1220309 [PubMed: 27759442]

Epstein RM (2007). Assessment in medical education. New England Journal of Medicine, 356(4), 387–396. 10.1056/NEJMra054784 [PubMed: 17251535]

Feingold A. (2013). A regression framework for effect size assessments in longitudinal modeling of group differences. Review of General Psychology, 17(1), 111–121. 10.1037/a0030048 [PubMed: 23956615]

Garland AF, Hurlburt MS, & Hawley KM (2006). Examining psychotherapy processes in a services research context. Clinical Psychology: Science and Practice, 13(1), 30–46.

Glasgow RE, & Riley WT (2013). Pragmatic measures: What they are and why we need them. American Journal of Preventive Medicine, 45(2), 237–243. 10.1016/j.amepre.2013.03.010 [PubMed: 23867032]

Guerra CE, Jacobs SE, Holmes JH, & Shea JA (2007). Are physicians discussing prostate cancer screening with their patients and why or why not? A pilot study. Journal of General Internal Medicine, 22(7), 901–907. 10.1007/s11606-007-0142-3 [PubMed: 17549576]

Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, & Conde JG (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. Journal of Biomedical Informatics, 42(2), 377–381. 10.1016/j.jbi.2008.08.010 [PubMed: 18929686]

Higa-McMillan CK, Francis SE, Rith-Najarian L, & Chorpita BF (2016). Evidence base update: 50 years of research on treatment for child and adolescent anxiety. Journal of Clinical Child and Adolescent Psychology, 45(2), 91–113. 10.1080/15374416.2015.1046177 [PubMed: 26087438]

Hoagwood KE, Olin SS, Horwitz S, McKay M, Cleek A, Gleacher A, Lewandowski E, Nadeem E, Acri M, & Chor KHB (2014). Scaling up evidence-based practices for children and families in New York state: toward evidence-based policies on implementation for state mental health systems. Journal of Clinical Child and Adolescent Psychology, 43(2), 145–157. 10.1080/15374416.2013.869749 [PubMed: 24460518]

Hofmann SG, Asnaani A, Vonk IJ, Sawyer AT, & Fang A. (2012). The efficacy of cognitive behavioral therapy: A review of meta-analyses. Cognitive Therapy and Research, 36(5), 427–440. 10.1007/s10608-012-9476-1 [PubMed: 23459093]

Hogue A, Dauber S, Lichvar E, Bobek M, & Henderson CE (2015). Validity of therapist self-report ratings of fidelity to evidence-based practices for adolescent behavior problems: Correspondence between therapists and observers. Administration and Policy in Mental Health and Mental Health Services Research, 42(2), 229–243. 10.1007/s10488-014-0548-2 [PubMed: 24711046]

Hogue A, Henderson CE, Dauber S, Barajas PC, Fried A, & Liddle HA (2008). Treatment adherence, competence, and outcome in individual and family therapy for adolescent behavior problems. Journal of Consulting and Clinical Psychology, 76(4), 544–555. 10.1037/0022-006X.76.4.544 [PubMed: 18665684]

Huey SJ Jr, Henggeler SW, Brondino MJ, & Pickrel SG (2000). Mechanisms of change in multisystemic therapy: Reducing delinquent behavior through therapist adherence and improved family and peer functioning. Journal of Consulting and Clinical Psychology, 68(3), 451–467. 10.1037/0022-006X.68.3.451 [PubMed: 10883562]

Hurlburt MS, Garland AF, Nguyen K, & Brookman-Frazee L. (2010). Child and family therapy process: concordance of therapist and observational perspectives. Administration and Policy in Mental Health and Mental Health Services Research, 37(3), 230–244. 10.1007/s10488-009-0251-x [PubMed: 19902347]

Killeen TK, Brady KT, Gold PB, Tyson C, & Simpson KN (2004). Comparison of self-report versus agency records of service utilization in a community sample of individuals with alcohol use disorders. Drug and Alcohol Dependence, 73(2), 141–147. 10.1016/j.drugalcdep.2003.09.006 [PubMed: 14725953]

Lewis CC, Fischer S, Weiner BJ, Stanick C, Kim M, & Martinez RG (2015). Outcomes for implementation science: an enhanced systematic review of instruments using evidence-based rating criteria. Implementation Science, 10(1), 155. 10.1186/s13012-015-0342-x [PubMed: 26537706]

Little RJ (1988). A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association, 83(404), 1198–1202. 10.1080/01621459.1988.10478722

Martino S, Ball S, Nich C, Frankforter TL, & Carroll KM (2009). Correspondence of motivational enhancement treatment integrity ratings among therapists, supervisors, and observers. Psychotherapy Research, 19(2), 181–193. 10.1080/10503300802688460 [PubMed: 19396649]

McLeod BD, Smith MM, Southam-Gerow MA, Weisz JR, & Kendall PC (2015). Measuring treatment differentiation for implementation research: The Therapy Process Observational Coding System for Child Psychotherapy Revised Strategies Scale. Psychological Assessment, 27(1), 314–325. 10.1037/pas0000037 [PubMed: 25346995]

McLeod BD, Southam-Gerow MA, Jensen-Doss A, Hogue A, Kendall PC, & Weisz JR (2019). Benchmarking treatment adherence and therapist competence in individual cognitive-behavioral treatment for youth anxiety disorders. Journal of Clinical Child & Adolescent Psychology, 48(sup1), S234–S246. 10.1080/15374416.2017.1381914 [PubMed: 29053382]

McLeod BD, Southam‐Gerow MA, Tully CB, Rodríguez A, & Smith MM (2013). Making a case for treatment integrity as a psychosocial treatment quality indicator for youth mental health care. Clinical Psychology: Science and Practice, 20(1), 14–32. 10.1111/cpsp.12020 [PubMed: 23935254]

McLeod BD, & Weisz JR (2010). The Therapy Process Observational Coding System for Child Psychotherapy Strategies Scale. Journal of Clinical Child and Adolescent Psychology, 39(3), 436–443. 10.1080/15374411003691750 [PubMed: 20419583]

Miller WR, & Rollnick S. (2014). The effectiveness and ineffectiveness of complex behavioral interventions: Impact of treatment fidelity. Contemporary Clinical Trials, 37(2), 234–241. 10.1016/j.cct.2014.01.005 [PubMed: 24469237]

Mufson L, & Sills R. (2006). Interpersonal psychotherapy for depressed adolescents (IPT-A): An overview. Nordic Journal of Psychiatry, 60(6), 431–437. 10.1080/08039480601022397 [PubMed: 17162450]

Perepletchikova F, Treat TA, & Kazdin AE (2007). Treatment integrity in psychotherapy research: analysis of the studies and examination of the associated factors. Journal of Consulting and Clinical psychology, 75(6), 829. 10.1037/0022-006X.75.6.829 [PubMed: 18085901]

Powell BJ, Beidas RS, Rubin RM, Stewart RE, Wolk CB, Matlin SL, … & Mandell DS (2016). Applying the policy ecology framework to Philadelphia's behavioral health transformation efforts. Administration and Policy in Mental Health and Mental Health Services Research, 43(6), 909–926. 10.1007/s10488-016-0733-6 [PubMed: 27032411]

Proctor EK, Silmere H, Raghavan R, Hovmand P, Aarons GA, Bunger A, Griffey R, & Hensley M. (2011). Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. Administration and Policy in Mental Health and Mental Health Services Research, 38(2), 65–76. 10.1007/s10488-010-0319-7 [PubMed: 20957426]

Rapley HA, & Loades ME (2019). A systematic review exploring therapist competence, adherence, and therapy outcomes in individual CBT for children and young people. Psychotherapy Research, 29(8), 1010–1019. 10.1080/10503307.2018.1464681 [PubMed: 29683046]

Regan J, Lau AS, Barnett M, Stadnick N, Hamilton A, Pesanti K, Bando L, & Brookman-Frazee L. (2017). Agency responses to a system-driven implementation of multiple evidence-based practices in children's mental health services. BMC Health Services Research, 17(1), 671. 10.1186/s12913-017-2613-5 [PubMed: 28927407]

Rodriguez-Quintana N, Lewis CC (2018). Observational coding training methods for CBT treatment fidelity: A systematic review. Cognitive Therapy and Research, 42(4), 358–368. 10.1007/s10608-018-9898-5

Schoenwald SK, Garland AF, Chapman JE, Frazier SL, Sheidow AJ, & Southam-Gerow MA (2011). Toward the effective and efficient measurement of implementation fidelity. Administration and Policy in Mental Health and Mental Health Services Research, 38(1), 32–43. 10.1007/s10488-010-0321-0 [PubMed: 20957425]

Simons AD, Padesky CA, Montemarano J, Lewis CC, Murakami J, Lamb K, DeVinney S, Reid M, Smith DA, & Beck AT (2010). Training and dissemination of cognitive behavior therapy for depression in adults: A preliminary examination of therapist competence and client outcomes. Journal of Consulting and Clinical Psychology, 78(5), 751–756. 10.1037/a0020569 [PubMed: 20873911]

Smith MM, McLeod BD, Southam-Gerow MA, Jensen-Doss A, Kendall PC, & Weisz JR (2017). Does the delivery of CBT for youth anxiety differ across research and practice settings? Behavior Therapy, 48(4), 501–516. 10.1016/j.beth.2016.07.004 [PubMed: 28577586]

Southam-Gerow MA, Chapman JE, Martinez RG, McLeod BD, Hogue A, Weisz JR, & Kendall PC (2021). Are therapist adherence and competence related to clinical outcomes in cognitive-behavioral treatment for youth anxiety?. Journal of Consulting and Clinical Psychology, 89(3), 188. [PubMed: 33829807]

Substance Abuse and Mental Health Services Administration. (2019). National Mental Health Services Survey (N-MHSS): 2018. Data on Mental Health Treatment Facilities. https://www.samhsa.gov/data/data-we-collect/nmhss-national-mental-health-services-survey

Thomas R, Abell B, Webb HJ, Avdagic E, & Zimmer-Gembeck MJ (2017). Parent-child interaction therapy: a meta-analysis. Pediatrics, 140(3). 10.1542/peds.2017-0352

Weisz JR, Kuppens S, Ng MY, Eckshtain D, Ugueto AM, Vaughn-Coaxum R, Jensen-Doss A, Hawley KM, Krumholz Marchette LS, & Chu BC (2017). What five decades of research tells us about the effects of youth psychological therapy: A multilevel meta-analysis and implications for science and practice. American Psychologist, 72(2), 79–117. 10.1037/a0040360 [PubMed: 28221063]

Young J, & Beck AT (1980). Cognitive therapy scale: Rating manual. Unpublished Manuscrip. https://beckinstitute.org/

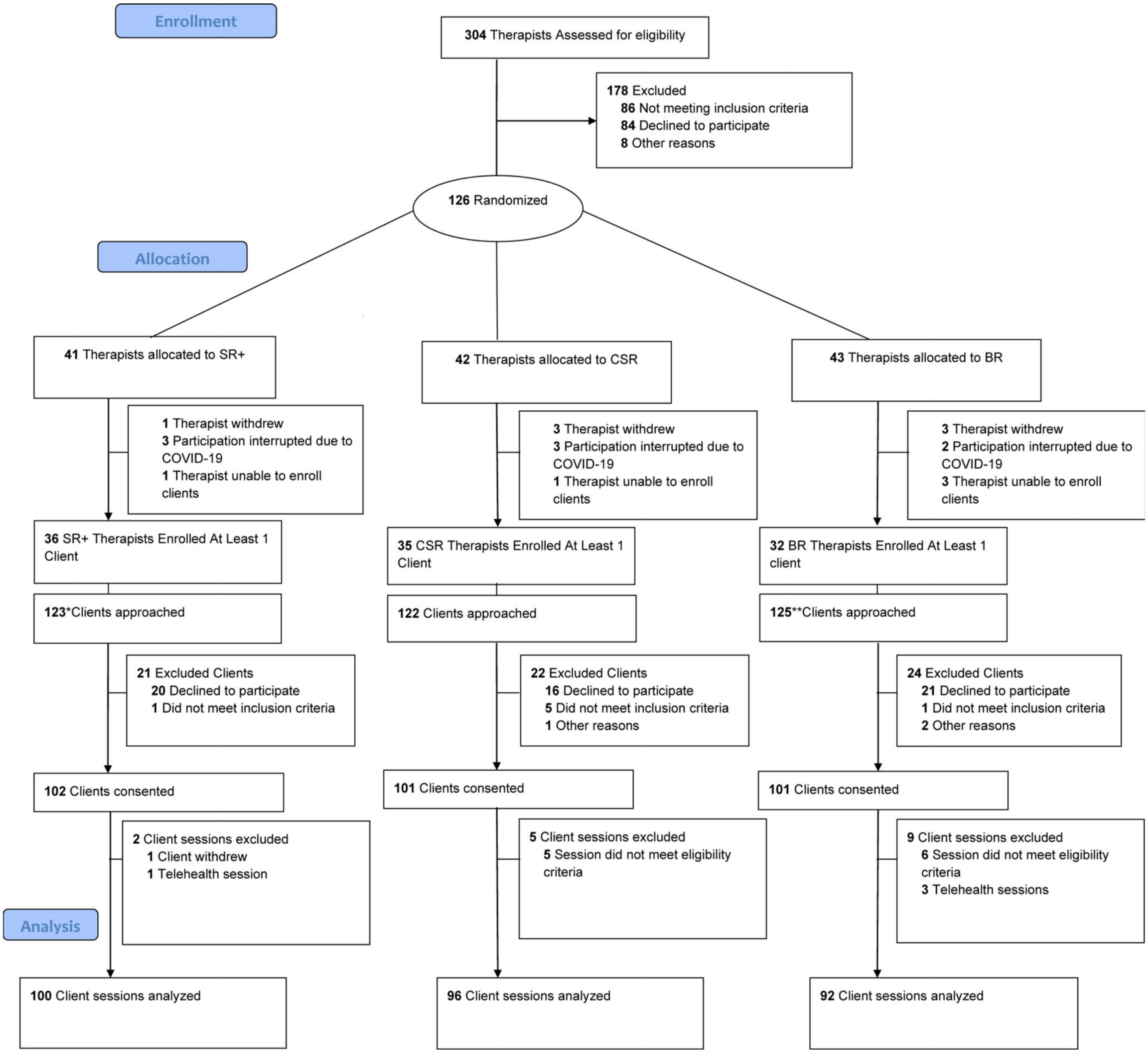
Enrollment
304 Therapists Assessed for eligibility
178 Excluded
86 Not meeting inclusion criteria
84 Declined to participate
8 Other reasons
126 Randomized
Allocation
41 Therapists allocated to SR+
42 Therapists allocated to CSR
43 Therapists allocated to BR
1 Therapist withdrew
3 Participation interrupted due to COVID-19
1 Therapist unable to enroll clients
3 Therapist withdrew
3 Participation interrupted due to COVID-19
1 Therapist unable to enroll clients
3 Therapist withdrew
2 Participation interrupted due to COVID-19
3 Therapist unable to enroll clients
36 SR+ Therapists Enrolled At Least 1 Client
35 CSR Therapists Enrolled At Least 1 Client
32 BR Therapists Enrolled At Least 1 client
123*Clients approached
122 Clients approached
125**Clients approached
21 Excluded Clients
20 Declined to participate
1 Did not meet inclusion criteria
22 Excluded Clients
16 Declined to participate
5 Did not meet inclusion criteria
1 Other reasons
24 Excluded Clients
21 Declined to participate
1 Did not meet inclusion criteria
2 Other reasons
102 Clients consented
101 Clients consented
101 Clients consented
2 Client sessions excluded
1 Client withdrew
1 Telehealth session
5 Client sessions excluded
5 Session did not meet eligibility criteria
9 Client sessions excluded
6 Session did not meet eligibility criteria
3 Telehealth sessions
Analysis
100 Client sessions analyzed
96 Client sessions analyzed
92 Client sessions analyzed

Figure 1.

CONSORT Flow Diagram

Note. To be eligible a therapist had to have already had a least one previous session with the client, the recorded session had to be in English, and therapists had to self-report doing at least 10 minutes of CBT with their client. Additionally, therapists in the Chart Stimulated Recall condition had to complete their chart note prior to conducting the chart stimulated recall for session to be eligible. The behavioral rehearsal condition was changed significantly after the pilot and thus those pilot sessions were excluded from analyses. *123 clients were approached from 37 therapists. One therapist was unable to continue participation in the study as they only had three eligible clients all of whom declined to participate in the study. **125 clients were approached from 35 therapists. One therapist was unable to continue participation in the study as they only had three eligible clients all of whom declined to participate in the study. One therapist was excluded because we were piloting our condition method. One therapist was excluded from analysis because all sessions recorded with the therapist were telehealth sessions due to COVID-19.

Table 1.

Therapist demographics by condition

	Condition			Test Statistic	p value
	SR *n* = 36	**BR *n* = 32**	**CSR *n* = 35**		
Age, *M (SD)*	38.80 (13.83)	36.35 (12.65)	37.77 (12.25)	$F_{(2, 98)} = 0.295$	0.745
Female Gender, *n* (%)	23 (64.9)	28 (87.5)	27 (77.1)	$\chi^2_{(2)} = 5.20$	0.074
Hispanic/Latinx, *n* (%)	0 (0)	4 (12.5)	2 (5.7)	$\chi^2_{(2)} = 4.62$	0.100
Racial Identify[a]					
White/Caucasian, *n* (%)	23 (64.9)	22 (68.8)	29 (82.9)	$\chi^2_{(2)} = 1.32$	0.516
Asian or Pacific Islander, *n* (%)	0 (0)	3 (9.4)	2 (5.7)	$\chi^2_{(2)} = 3.20;$	0.202
Black or African American, *n* (%)	9 (25.0)	5 (15.6)	4 (11.4)	$\chi^2_{(2)} = 3.19$	0.203
Other, *n* (%)	1	1 (3.1)	0	$\chi^2_{(2)} =1.30$	0.523
Education				$\chi^2_{(4)} =3.51$	.476
Bachelor’s Degree	0 (0)	1 (3.1)	1 (2.9)		
Master’s Degree	36 (100)	29 (90.6)	33 (94.3)		
Doctorate	0 (0)	2 (6.3)	1 (2.9)		
Years experience, *M (SD)*	11.44 (10.96)	9.28 (10.95)	10.29 (10.95)	$F_{(2, 100)}= 0.36$	.702
Type of employee, *n* (%)				$\chi^2_{(2)} = .297$	.862
Salaried	19 (52.8)	17 (53.1)	17 (48.6)		
Fee for service	16 (44.4)	13 (40.6)	17 (48.6)		
Participated in city sponsored Evidence-Based Practice training initiative, *n* (%)	25 (69.4)	20 (62.5)	23 (65.7)	$\chi^2_{(2)} =0.37$	.833
Number of client sessions recorded in the study, *M (SD)*	2.75 (.55)	2.90 (.39)	2.74 (.61)	$F_{(2, 100)} = 1.00$	.371

Abbreviations: SR, Self-Report; CSR, Chart-Stimulated Recall; BR, Behavioral Rehearsal.

[a] Numbers do not add to 100% as respondents could endorse identification with more than one race; in addition, 12 (12.0%), 12 (12.5%), and 10 (10.9%) did not disclose their race within the SR, CSR, and BR conditions, respectively.

Table 2.

Client demographics by condition

	Condition			Test Statistic	p value
	SR *n* = 100	**BR *n* = 93**	**CSR *n* = 96**		
Age, *M (SD)*	13.34 (3.79)	13.41 (3.97)	13.42 (3.92)	$F_{(2,\ 284)}$ = .01	.987
Female Gender, *n* (%)	40 (40.0)	43 (46.3)	37 (38.5)	$\chi^2_{(2)}$ =2.14	.343
Hispanic/Latinx, *n* (%)	27 (27.0)	16 (17.2)	21(21.9)	$\chi^2_{(2)}$ =2.24	.326
Racial Identify					
White/Caucasian, *n* (%)	36 (36.0)	34 (36.6)	36 (37.5)	$\chi^2_{(2)}$ = .05	.976
Asian or Pacific Islander, *n* (%)	3 (3.0)	5 (5.4)	1(1.0)	$\chi^2_{(2)}$ =2.64	.267
Black or African American, *n* (%)	50 (50.0)	46 (49.5)	43 (44.8)	$\chi^2_{(2)}$ =.70	.706
Other, *n* (%)	6 (6.0)	9 (9.7)	6 (6.2)	$\chi^2_{(2)}$ =1.25	.537
Number of previous treatment sessions with therapist prior to study participation, *M (SD)*	29.10 (50.98)	31.97 (47.83)	21.99 (27.25)	$F_{(2,\ 283)}$ = 1.38	.254

Abbreviations: SR, Self-Report; CSR, Chart-Stimulated Recall; BR, Behavioral Rehearsal.

[a] Numbers do not add to 100% as respondents could endorse identification with more than one race; in addition, 4 (11.1%), 0 (0.0%), and 1 (3.1%) did not disclose their race within the SR, CSR, and BR conditions, respectively.

Table 3.

Comparison of Fidelity Condition Scores to Direct Observation Scores

	Mean Condition Score (SD)	**Mean direct observation Score (SD)**	**Least Squares Mean Paired Difference Between Condition and direct observation**	**Cohen's d**[d]	**P value of Paired Difference**[b]
Maximum CBT Score					
SR (n=100)	5.28 (1.17)	3.42 (1.55)	−1.84	1.34	<.001
CSR (n=96)	4.53 (1.18)	3.69 (1.36)	−0.83	0.62	<.001
BR (n=92)	4.09 (1.48)	3.72 (1.59)	−0.34	0.22	.08[c]
Mean CBT Score[a]					
SR (n= 89)	3.91 (0.82)	2.77 (0.72)	−1.15	1.49	<.001
CSR (n= 94)	3.18 (0.63)	2.80 (0.68)	−0.38	0.58	.002
BR (n=87)	3.13 (0.93)	2.82 (0.70)	−0.30	0.36	.02[c]
CBT Count Total Techniques Scored					
SR (n=100)	7.17 (2.59)	3.15 (1.84)	−4.02	1.79	<.001
CSR (n=96)	4.38 (1.74)	3.15 (1.40)	−1.23	0.77	<.001
BR (n=92)	3.10 (1.58)	3.21 (1.78)	0.118	0.07	.75[c]

Abbreviations: SR, Self-Report; CSR, Chart-Stimulated Recall; BR, Behavioral Rehearsal. Direct Observation was measured by the Therapy Process Observation Coding Scale-Revised Strategies.

[a] Sample sizes varied for this analysis, as it only included sessions in which CBT was scored as present (i.e., a 2 or greater on the 7-point scale).

[b] Presents the *p* value of the significance of the intercept of the three-level regression model comparing condition and direct observation adherence score. A significant *p* value indicates that the least squares paired mean difference is not equal to zero.

[c] $p > .01$, indicating no significant difference in adherence score produced between Condition (SR, CSR, or BR) and direct observation.

[d] Cohen's d calculated as the magnitude of effect of the least squares mean paired difference between condition and direct observation

Table 4.

Relative Performance of Each Fidelity Condition

	Difference in LS Paired Mean Difference between Conditions	*p* value[b]
Maximum CBT Score		
SR vs. CSR	1.01	<.001[***]
SR vs. BR	1.50	<.001[***]
CSR vs. BR	0.49	.15
Mean CBT Score[a]		
SR vs. CSR	0.77	<.001[***]
SR vs. BR	0.85	<.001[***]
CSR vs. BR	−0.08	.90
CBT Count Total Techniques Scored		
SR vs. CSR	2.79	.002[**]
SR vs. BR	4.14	<.001[***]
CSR vs. BR	1.35	.01[*]

Abbreviations: SR, Self-Report; CSR, Chart Stimulated Recall; BR, Behavioral Rehearsal.

[a] Sample sizes varied for this analysis, as it only included sessions in which CBT was scored as present (i.e., a 2 or greater on the 7-point scale).

[b] *p* value tests whether the paired mean difference is different between conditions; includes a Tukey-Kramer adjustment. Significant *p* values indicate that conditions differ in their relative performance.

[*] $p < .05$.

[**] $p < .01$.

[***] $p < .001$.