Administration and Policy in Mental Health and Mental Health Services Research
https://doi.org/10.1007/s10488-023-01274-x

ORIGINAL ARTICLE

Development and Evaluation of a Pragmatic Measure of Adherence to Dialectical Behavior Therapy: The DBT Adherence Checklist for Individual Therapy

Melanie S. Harned[1,2,3] · Sara C. Schmidt[1] · Kathryn E. Korslund[4] · Robert J. Gallop[5]

Accepted: 8 May 2023

Abstract

This paper presents two studies conducted to develop and evaluate a new pragmatic measure of therapist adherence to Dialectical Behavior Therapy (DBT): the DBT Adherence Checklist for Individual Therapy (DBT AC-I). Study 1 used item response analysis to select items from the gold standard DBT Adherence Coding Scale (DBT ACS) using archival data from 1271 DBT sessions. Items were then iteratively refined based on feedback from 33 target end-users to ensure relevance, usability, and understandability. Study 2 examined the psychometric properties of the DBT AC-I as a therapist self-report and observer-rated measure in 100 sessions from 50 therapist-client dyads, while also evaluating predictors of therapist accuracy in self-rated adherence. When used as a therapist self-report measure, concordance between therapist and observer ratings was at least moderate ($AC_1 \geq 0.41$) for all DBT AC-I items but overall concordance (ICC = 0.09) as well as convergent ($r = 0.05$) and criterion validity (AUC = 0.54) with the DBT ACS were poor. Higher therapist accuracy was predicted by greater DBT knowledge and adherence as well as more severe client suicidal ideation. When used by trained observers, the DBT AC-I had excellent interrater reliability (ICC = 0.93), convergent validity ($r = 0.90$), and criterion validity (AUC = 0.94). While therapists' self-rated adherence on the DBT AC-I should not be assumed to reflect their actual adherence, some therapists may self-rate accurately. The DBT AC-I offers an effective and relatively efficient method of evaluating adherence to DBT when used by trained observers.

Keywords Adherence · Fidelity · Dialectical behavior therapy · Psychometric

Dialectical Behavior Therapy (DBT; Linehan, 1993) is an evidence-based psychotherapy (EBP) that is primarily used to treat borderline personality disorder (BPD; Storebø et al., 2020) and self-injurious behaviors (SIB; DeCou et al., 2019). Substantial efforts have been made to implement DBT around the world at the clinic, system, state, and national levels (e.g., Carmel et al., 2014; DuBose et al., 2019; Flynn et al., 2020; Herschell et al., 2014). Encouragingly, research has found that leading DBT training models result in high rates of adoption and sustainability of DBT in diverse healthcare settings (e.g., Navarro-Haro et al., 2019; Swales et al., 2012). However, little is known about the degree to which DBT is delivered with fidelity once adopted, which is a critical aspect of quality assurance and indicator of implementation success (Proctor et al., 2009). Therapist adherence is a core component of fidelity and refers to the degree to which therapists deliver the treatment procedures as intended (Perepletchikova et al., 2007). In DBT, higher therapist adherence leads to better outcomes for clients (Harned et al., 2022), making it a particularly important target for implementation and quality assurance efforts.

A primary barrier to evaluating the impact of DBT implementation efforts on therapist adherence has been the lack of a pragmatic method of measuring adherence. As is typical

✉ Melanie S. Harned
melanie.harned@va.gov

[1] VA Puget Sound Health Care System, 1660 South Columbian Way, Mailstop S-116-MHC, Seattle, WA 98108, USA

[2] Department of Psychiatry & Behavioral Sciences, University of Washington, Seattle, WA, USA

[3] Department of Psychology, University of Washington, Seattle, WA, USA

[4] THIRA Health, Bellevue, WA, USA

[5] Department of Mathematics, West Chester University, West Chester, PA, USA

Published online: 07 June 2023

for adherence measures more broadly (Schoenwald & Garland, 2013), the gold standard measure of adherence to DBT utilizes direct observational methods in which trained coders rate recorded therapy sessions. The DBT Adherence Coding Scale (DBT ACS; Linehan & Korslund, 2003) is a comprehensive, observational measure that assesses the degree to which therapists sufficiently deliver each DBT strategy in a session. The DBT ACS has excellent inter-rater reliability, strong discriminant validity, and generates dependable adherence scores (Harned et al., 2021a, b). The DBT ACS is also effective in predicting reductions in key client outcomes, including suicide attempts, psychiatric hospitalizations, and treatment dropout (Harned et al., 2022).

Although the DBT ACS is a highly effective method of measuring adherence to DBT, it is not practical for use as a quality assurance tool in routine care. As is true with any observational adherence measure (Schoenwald et al., 2011), the DBT ACS requires considerable resources to use (e.g., time, money, equipment) and can be burdensome for both therapists and clients (e.g., due to having to record sessions). Additional barriers specific to the DBT ACS include its length and complexity (66 items coded on a 6-point Likert scale), the extensive training required to become a reliable coder, access restrictions that limit its availability to approved coders, and the high cost associated with its use. As a result, the DBT ACS has only been used for high stakes purposes, including treatment outcome research and therapist certification. This has left therapists, programs, and systems that have invested in the implementation of DBT with no way to evaluate if the treatment is being delivered in a manner consistent with the evidence-based intervention. This is particularly concerning given a recent study finding a high rate of non-adherent sessions (48%) among therapists trained in a system-level DBT initiative (Harned et al., 2021a, b). Thus, there is a critical need to develop pragmatic measures of DBT adherence that can be used to guide quality assurance efforts in diverse practice contexts.

Developing pragmatic measures of adherence is challenging because of the tension that exists between effectiveness (psychometric rigor) and efficiency (feasibility for use in routine care), both of which are difficult to achieve in a single instrument (Schoenwald et al., 2011). To date, most efforts to resolve this tension have focused on developing self-report measures for therapists to rate their own delivery of a treatment (Schoenwald & Garland, 2013). While therapist self-report measures are typically feasible (e.g., quick, inexpensive, nonintrusive), most studies have found low concordance between therapist self-report and observer ratings due to therapists tending to overestimate their adherence (e.g., Brookman-Frazee et al., 2021; Brosan et al., 2008; Hogue et al., 2015; Hurlburt et al., 2010; Martino et al., 2009). However, higher concordance has been found for some EBPs (e.g., Brookman-Frazee et al., 2021; Hogue et al., 2015) and when rating the presence/absence of techniques rather than frequency or competence (e.g., Martino et al., 2009). In addition, certain therapist characteristics (e.g., higher confidence, less professional experience) and session factors (e.g., ability to carry out planned activities) may improve the accuracy of self-ratings (Brookman-Frazee et al., 2021; Loades & Myles, 2016). Taken together, these findings suggest that therapist self-report offers an efficient method of measuring adherence but requires evaluation to determine whether and under which conditions it may be effective.

Given the challenges associated with therapist self-report, another option for improving the efficiency of adherence measurement without sacrificing effectiveness is to simplify and streamline gold standard observational measures. Potential modifications to observational measures that may improve efficiency include converting them to a checklist format, simplifying the rating scale, and reducing the number of items (Schoenwald et al., 2011). For the DBT ACS, this could involve condensing the measure to the most critical DBT strategies, collapsing the 6-point rating scale to a binary scale, and using a checklist format that includes behavioral anchors that describe adherent and non-adherent delivery of each strategy.

This paper reports the results of two studies that describe the development and evaluation of a new pragmatic measure of adherence to DBT: the DBT Adherence Checklist for Individual Therapy (DBT AC-I). Consistent with recommendations to use well-established observational measures as the foundation for developing pragmatic adherence measures (Schoenwald et al., 2011), the DBT AC-I was derived from the DBT ACS as this approach was expected to be more likely to yield a measure with strong psychometric properties and construct validity (Hogue et al., 2015). Overall, the aim of this project was to develop a measure that: (a) was brief and easy to use, (b) could be completed as a therapist self-report or observer-rated measure, (c) had strong psychometric properties, (d) could be used for quality assurance purposes in routine practice, and (e) would be free and publicly available. Study 1 focused on developing the DBT AC-I and an accompanying training manual based on item response analysis of archival DBT ACS data followed by iterative feedback from target end-users. Study 2 evaluated: (1) the reliability and validity of the DBT AC-I when used as a therapist self-report measure, (2) therapist-, client-, and session-level predictors of therapist accuracy in rating their own adherence on the DBT AC-I, and (3) the reliability and validity of the DBT AC-I when used by trained observers.

Study 1: Development of the DBT AC-I

Method

Participants

DBT Therapists ($n = 27$) Therapists were recruited from July to October 2019 by posting information about the study on listservs (email groups) for mental health professionals, including a DBT-specific listserv. Interested individuals were directed to complete an online screening questionnaire to determine eligibility. Inclusion criteria were: (1) currently delivering DBT individual therapy in a routine practice setting (defined as any setting other than an academic medical or research setting); (2) completed at least some training in DBT (defined as at least one of the following: 2 + days of workshops, a graduate level course in DBT, 6 + hours of online training, or reading any of Linehan's DBT manuals); (3) licensed or working under the supervision of a licensed mental health provider; and (4) age 18 + . Individuals were excluded if they: (1) had insufficient English proficiency; (2) did not reside in the United States; and/or (3) conducted professional training in DBT and/or were trained in the DBT ACS. See Table 1 for therapist characteristics.

Overall, 215 individuals completed the online therapist eligibility screen, 136 (63.3%) of whom reported meeting the eligibility criteria. Reasons for exclusion were: (1) did not reside in the United States ($n = 16$), (2) worked in an academic medical or research setting ($n = 9$), (3) did not provide DBT individual therapy ($n = 4$), (4) was unlicensed without a supervisor ($n = 2$), and (5) conducted professional training in DBT and/or was trained in the DBT ACS ($n = 48$). For each iteration of the formative evaluation, 6–12 unique therapists were recruited from the pool of eligible therapists with the goal of achieving a sample with varying degrees, practice settings, and levels of training in DBT. Of the 136 eligible therapists, 39 were invited to participate of whom 27 (69.2%) enrolled in the study. The remainder did not respond to recruitment emails ($n = 9$), declined participation ($n = 2$), or were no longer eligible to participate ($n = 1$).

DBT Experts ($n = 6$) DBT experts were recruited from the professional networks of the research team. Experts could participate in multiple iterations of the formative evaluation. Inclusion criteria were: (1) conducts professional training in DBT for a recognized DBT training company, and/or (2) is a reliable coder on the DBT ACS.

Table 1 Therapist characteristics

	Study 1 Therapists ($n = 27$)	Study 2 Therapists ($n = 50$)
Age, $M \pm SD$	37.3 ± 7.8	40.7 ± 11.7
Gender		
Female	21 (77.8)	42 (84.0)
Male	6 (22.2)	8 (16.0)
Race/Ethnicity		
Non-Hispanic White	25 (96.2)	42 (84.0)
Hispanic White	1 (3.7)	5 (10.0)
Asian/Asian-American	1 (3.7)	3 (6.0)
Degree		
Masters	17 (63.0)	36 (72.0)
Doctoral	10 (37.0)	14 (28.0)
Discipline		
Psychologist	10 (37.0)	12 (24.0)
Social worker	9 (33.3)	17 (34.0)
Mental health counselor	8 (29.6)	18 (36.0)
Other	0 (0.0)	3 (6.0)
DBT Program Setting		
Community mental health center	9 (33.3)	15 (30.0)
Private practice	12 (44.4)	31 (62.0)
Medical center	1 (3.7)	2 (4.0)
College counseling center	2 (7.4)	0 (0.0)
Residential/day treatment	1 (3.7)	2 (4.0)
Inpatient psychiatric unit	2 (7.4)	0 (0.0)
DBT Training		
Read Linehan's, 1993 DBT manual	27 (100.0)	50 (100.0)
Read Linehan's 2015 DBT skills manual	26 (96.3)	49 (98.0)
Graduate coursework/practicum	12 (44.4)	19 (38.0)
Supervision within agency	–	31 (62.0)
Supervision with an external expert	–	30 (60.0)
DBT online training course	13 (48.1)	25 (50.0)
DBT workshop	26 (96.3)	49 (98.0)
Days of workshop training, $M \pm SD$	17.8 ± 26.0	18.9 ± 16.4
DBT Experience		
# of years providing DBT, $M \pm SD$	–	6.0 ± 3.8
Very comfortable using DBT	–	36 (72.0)

Data are presented as n (%) unless otherwise noted

Procedures

All study procedures were approved by the VA Puget Sound Institutional Review Board.

Initial Scale Development

Original Measure Items were derived from the DBT ACS (Linehan & Korslund, 2003) that includes 66 items across 12 subscales that reflect the strategies used in DBT as specified in the treatment manual (Linehan, 1993). Not all strategies (items) are required in each session, and the measure includes "if/then" rules for when they must be used. Each item is rated on a scale where 0 = not used/not necessary, 1–3 = below adherence, 4 = minimum threshold for adherence, and 5 = adherent with high sufficiency. The DBT ACS is scored by averaging all non-zero items to create a computed global score where 4.0 and higher is considered "adherent."

Data Sources An archival dataset of DBT ACS data from six DBT clinical trials was used for analyses. Across the six trials, trained observers used the DBT ACS to rate 84 therapists during 1,271 DBT individual therapy sessions with 292 clients. In addition, 78 sessions were coded by the gold standard rater (Dr. K. Korslund) for the purpose of evaluating interrater reliability according to the procedures of each trial. The dataset included both efficacy and effectiveness trials of DBT with adults and adolescents from a variety of client populations. A detailed description of this archival dataset can be found in Harned et al., (2021a, b).

Item Selection The first step in reducing the pool of DBT ACS items was to remove items with poor pooled interrater reliability (IRR < 0.60; Cicchetti, 1994) and limited variability (Hinkin, 1998). The remaining items were then analyzed using item response theory (IRT) to identify the items that were most informative regarding the DBT ACS computed global score. IRT was selected because it improves upon classical test theory and is widely used in clinical research to shorten measures (Reise & Waller, 2009). Item response models are built around the assumption of unidimensionality. Given that the DBT ACS has multiple subscales, to accommodate unidimensionality we conducted exploratory and confirmatory factor analyses (EFA and CFA) on the remaining items. The sample was randomly split into two near equal sized sub-samples. EFA was conducted on the first subsample (i.e., developmental sample) to identify the factor structures and CFA was implemented on the second subsample (i.e., validation sample) to confirm the factor structures. Items were identified as loading within a factor (i.e., loadings ≥ 0.30) or non-loaders. We then fit separate IRT graded response models (GRMs) for all the loading items within each factor as well as for the collection of non-loading items. Unidimensionality was assessed based on the eigenvalues of polychoric correlations. Candidate items for inclusion were selected based on: (1) the slope parameters using a threshold of > 0.5 (An & Yung, 2014) to indicate the item was informative regarding overall adherence, and (2) item information curves (IIC), test information curves (TIC), and item characteristic curves (ICC) to determine how well the item discriminated between therapists at different levels of adherence. When these fit indices were comparable among items, we selected items that were viewed as important in defining DBT and maximized coverage of content across the 12 DBT ACS subscales. Feedback on the candidate items was obtained in the first phase of the formative evaluation (see below) before the items were finalized. A second IRT model was conducted with the final set of items to examine overall model fit as well as item slope and threshold parameters.

Rating Scale Selection Analyses were also conducted to evaluate whether the 6-level DBT ACS rating scale could be simplified to a binary scale. To that end, a total score was computed by dichotomizing each candidate item as non-adherent (0 = 1, 2, or 3) or adherent (1 = 0, 4, or 5) and summing them. To determine whether there was a significant loss in discrimination with the binary scale, the Area Under the Curve (AUC) was used to quantify the ability of this total score (sum of binary items) to correctly identify adherent versus non-adherent sessions according to the original DBT ACS computed global score. We set a benchmark for using the binary scale as achieving at least excellent discrimination (AUC ≥ 0.90) as well as being significantly correlated with the DBT ACS computed global score.

Formative Evaluation

Consistent with best practices from product development models, the DBT AC-I and the accompanying manual were tested and iteratively revised in a formative evaluation. Revisions were based on feedback from the target audience (DBT clinicians) and DBT experts with the goal of ensuring relevance, usability, and understandability.

Expert Interviews In the first phase, feedback was obtained from DBT experts via individual interviews about the relevance and breadth of the candidate items (i.e., whether the items included the most critical strategies of DBT), item wording and clarity, and the rating scale. This expert feedback was used to finalize the items that were included in the measure. Experts were paid $40 to complete an interview.

Therapist Surveys In the second phase, DBT therapists provided feedback via an online survey that asked them to: (1) rate each item in terms of how easy it was to understand on a scale ranging from 1 (not at all) to 5 (extremely), and (2) suggest changes for each item. The measure was then iteratively revised based on therapist feedback until adequate understandability was achieved, which was defined as at least a mean rating of 4 ("very" easy to understand) out of 5 for each DBT AC-I item. Therapists were paid $30 to complete an online survey.

Think-Aloud Interviews The third phase involved conducting think-aloud interviews with DBT therapists and experts to evaluate the measure's usability and effectiveness in enabling participants to think like expert adherence coders. Think-aloud interviews are a method of capturing participant thought processes when engaging with an instrument (Wolcott & Lobczowski, 2021). Participants were instructed to complete the measure about a recent session of their own (therapists) or of a supervisee (experts) while continuously verbalizing their thought processes out loud (e.g., about why they decided to rate an item as adherent or non-adherent). During the think-aloud process, the interviewer transcribed participants' thoughts and did not speak except to prompt them to continue talking if they were silent for more than 20 s. Participants were then asked to provide qualitative feedback on the measure and the manual as well as to complete the 10-item System Usability Scale (SUS; Brooke, 1996), a widely used measure of usability that has excellent psychometric properties (Bangor et al., 2008). All SUS items were modified by changing the generic term "the system" to refer specifically to the DBT AC-I (e.g., "I thought the DBT Adherence Checklist was easy to use"). Items are rated on a 5-point scale from 1 (strongly disagree) to 5 (strongly agree) and used to generate a total score ranging from 0 (negative) to 100 (positive) where scores of 68 or higher indicate above average usability (Bangor et al., 2008). Therapists were paid $30 and experts were paid $40 to complete the think-aloud interviews.

Expert review and finalization After additional revisions were made, a final round of written feedback was obtained from DBT experts before the measure and manual were finalized for validation in Study 2. Experts were paid $40 to provide this final review.

Results

Initial Scale Development

Of the 66 DBT ACS items, 17 were removed due to having poor pooled interrater reliability (IRR < 0.60) and 4 were removed due to having extremely low variability (coefficient of variation < 5%). The remaining pool of 45 items was analyzed using IRT. Based on the IRT results, 24 items were initially selected as candidates for inclusion in the measure that showed strong overall model fit. Specifically, the 24 items explained 90.6% and 88.3% of the total information for the latent trait of adherence compared to the 45 items included in the IRT and all 66 DBT ACS items, respectively. To evaluate whether the continuous (0–5) rating scale could be collapsed to a binary (0–1) scale, a revised total score was created by dichotomizing and summing the 24 candidate items. Analyses indicated that this revised total score had excellent discrimination between adherent versus non-adherent sessions (AUC = 0.90, SE = 0.01) and was significantly correlated with the original DBT ACS computed global score (r = 0.70, p < 0.0001). Thus, a binary scale was selected for the measure to reduce complexity. Each candidate item was then written using a common format with behavioral anchors describing adherent (1) and non-adherent (0) use of the strategy based on the definitions provided in the DBT ACS.

Formative Evaluation

The initial 24-item measure was then iteratively refined based on therapist and expert feedback. In total, four iterations were completed that included expert interviews (n = 6), therapist surveys (n = 21), think-aloud interviews (n = 6 therapists and 3 experts), and final review by experts (n = 2). In this process, 4 items were added and 3 items were removed to improve construct and content validity, item wording was refined to enhance clarity, and the training manual was developed. The final version of the measure included 25 items rated on a binary scale that were each rated as "very" to "extremely" easy to understand (M's = 4.1–4.9 out of 5). In addition, the measure was rated as highly usable on the SUS (M = 80.0, SD = 8.6).

Fit of the Final Scale

Given that items were added and removed during the formative evaluation process, an IRT model was re-run on the final set of 25 items. Overall, the model showed a good fit with our data. Specifically, the 25 items explained 92.1 and 89.4% of the total information for the latent trait of adherence compared to the 45 items included in the IRT and all 66 DBT ACS items, respectively. All but one item had a slope parameter ≥ 0.5, indicating the items provided important information about the overall construct of adherence (see Supplemental Table 1 for all parameter estimates). The one item with a slope < 0.5 (informal exposure) was viewed as important from a content validity perspective and was therefore retained despite its low informativeness. To confirm the use of a binary rating scale, the AUC was re-computed using the revised total score (sum of the final 25 dichotomized items). This total score yielded excellent discrimination between adherent and non-adherent sessions (AUC = 0.91, SE = 0.01) and was highly correlated with the DBT ACS computed global score (r = 0.72, p < 0.0001).

Study 2: Psychometric Evaluation of the DBT AC-I

Method

Participants

Therapists ($n=50$) Therapists were recruited from January 2020 to February 2021 by posting information about the study on listservs for mental health professionals, including a DBT-specific listserv, emailing therapists who self-identified as delivering DBT in the Psychology Today online database, and emailing therapists who had previously attended DBT workshops conducted by Behavioral Tech, LLC. Interested individuals were directed to complete an online screening questionnaire to determine eligibility. Inclusion criteria were the same as in Study 1 plus therapists were required to: (1) be able to video-record and share therapy sessions with the study team and (2) have one DBT client (18+) who was able to attend treatment for the duration of the study period and consented to having their sessions recorded and shared with the study team. Individuals were excluded if they: (1) had insufficient English proficiency; (2) had already participated in Study 1; (3) were trained in the DBT ACS; and/or (4) did not reside in the United States. Eligible therapists who were invited to participate in the study were sent an Information Statement as well as a packet of information to provide to a study-eligible client.

Clients ($n=50$) Therapists were asked to recruit a study-eligible client from their caseload and to have interested clients complete a permission to contact form. Once therapists returned the signed permission to contact form, a member of the study team called the client to describe study procedures and answer any questions. Interested clients were sent an Information Statement and HIPAA authorization form to complete and return to the research team. Upon receipt of the signed consent materials, clients were compensated $30 for their time.

Procedures

All study procedures were approved by the VA Puget Sound Institutional Review Board.

Study Assessments Data collection occurred from February 2020 to July 2021 and involved having therapists complete four online surveys over an approximately 4-month period. Therapists were sent a link to access the baseline survey after their client returned their consent materials. Therapists who completed the baseline assessment were considered enrolled in the study. Upon enrollment, therapists were emailed electronic copies of the DBT AC-I and its accompanying manual and asked to review them prior to using the measure to rate their first session. Each therapist was also mailed an encrypted USB key on which to save and return their video-recorded therapy sessions to the research team. Therapists were asked to video-record the first two sessions with their study client that occurred following their baseline assessment and to complete a brief online survey after each session that included the DBT AC-I. Therapists were asked to complete the DBT AC-I as close to the end of the session as possible and were told they could initially complete it on paper if needed prior to entering it into the online survey. Therapists also completed a follow-up survey 3 months after the second session. Therapists were paid $30 for completion of the baseline survey, $30 for completion of each of the two session recordings and associated surveys, and $40 for completion of the follow up survey.

Participant Flow and Retention Overall, 253 individuals completed the online eligibility screen, 173 (68.4%) of whom met the eligibility criteria. Reasons for exclusion were: (1) did not reside in the United States ($n=29$), (2) participated in Study 1 ($n=16$), (3) worked in an academic medical or research setting ($n=8$), (4) did not have an eligible client ($n=8$), (5) did not provide DBT individual therapy ($n=3$), (6) was unlicensed without a supervisor ($n=1$), (7) was trained in the DBT ACS ($n=1$), and (8) was unable to video-record sessions ($n=1$). Thirteen individuals were deemed potentially eligible and in need of additional follow up if there were insufficient eligible therapists. Therapists were selected to participate based on the goal of achieving a sample with a variety of degrees, practice settings, and levels of training in DBT. Of the 173 eligible therapists, 133 were invited to participate of whom 53 (39.8%) enrolled in the study. The remainder did not respond to recruitment emails ($n=15$), failed to return completed permission to contact forms ($n=20$), declined participation ($n=23$), could not get agency approval to participate ($n=6$), did not have an eligible client ($n=6$), were unable to participate due to COVID pandemic-related changes to their work environment ($n=6$), or did not complete the baseline assessment ($n=4$). Of the 53 enrolled clinicians, 50 completed all the study procedures and are included in the analyses.

Observational Coding Raters ($n=3$) were selected for this study who had previously been trained to reliability on the DBT ACS at the level of the subscale by Dr. K. Korslund, the gold standard rater. Prior to commencing coding for the present study, all raters coded a calibration session to ensure they had not drifted from the gold standard rater. During the present study, a total of 100 sessions (2 per therapist/

client dyad) were coded using the DBT ACS. Two raters were assigned to code each therapist (1 per session) to minimize the risk of rater bias. In addition, reliability checks were periodically conducted during the study for each rater using a random selection of 10% of all coded sessions ($n = 10$). Reliability was evaluated by comparing the rater's scores to those of the gold standard rater. Interrater reliability was excellent for the DBT ACS computed global score (ICC = 0.98, 95% CI = 0.90–0.99).

Measures

Therapist-Level Measures

Therapist Characteristics At baseline, therapists provided information about their demographics, professional characteristics, DBT program, DBT training/experience, and self-efficacy in delivering DBT (ranging from 1 'not at all comfortable' to 4 'very comfortable').

DBT Knowledge A 46-item multiple choice test with 4 response choices per item was developed by study investigators to assess knowledge of DBT strategies and skills as well as ability to apply knowledge in hypothetical clinical scenarios. Items were developed to map onto the therapist strategies assessed in the DBT ACS and multiple iterations were piloted with DBT and non-DBT therapists during Study 1 before finalizing the measure for use in this study. The score used for analysis was the proportion of items correct. The test had excellent discriminant validity between DBT therapists in the present study ($n = 50$, 81.5% correct) and therapists with no training in DBT ($n = 49$, 35.4% correct), t (97) = 23.3, $p < 0.001$, $d = 4.68$.

Client-Level Measures

Client Demographics and DBT Treatment Therapists reported on their clients' demographics and current DBT treatment modes at baseline.

Psychiatric Functioning Therapists reported their clients' current psychiatric diagnoses at baseline based on DSM-5 diagnostic codes. Therapists also rated their clients' global functioning via the Global Assessment Scale (GAS; Endicott et al., 1976), a 0–100 scale of the overall severity of illness where higher scores indicate better functioning.

Self-Injurious Thoughts and Behaviors Therapists completed the informant report version of the Columbia Suicide Severity Rating Scale (C-SSRS; Posner et al., 2011) to assess the number of times their client attempted suicide (actual, aborted, or interrupted attempts) and engaged in non-suicidal self-injury (NSSI) in the 3 months prior to the baseline assessment. The C-SSRS also assessed the severity of their clients' suicidal ideation (SI) in the past 3 months ranging from 0 (none) to 5 (active SI with specific plan and intent).

Session-Level Measures

Therapist Self-Reported Adherence The 25-item DBT AC-I was used to assess therapists' self-reported adherence to DBT in two consecutive individual therapy sessions. A total score was calculated by summing all items (range = 0–25). In addition, therapists completed items assessing: (1) when they completed the DBT AC-I in relation to the session; (2) whether they read the DBT AC-I manual and/or referred to it while rating the session; (3) how long it took to complete the DBT AC-I, and (4) how helpful the DBT AC-I manual was as a training method (ranging from 0 'not at all' to 5 'extremely'). At the Session 2 assessment, therapists also completed the SUS (Brooke, 1996; see Study 1 for a description).

Observer-Rated Adherence The DBT ACS (Linehan & Korslund, 2003) was used to code each therapy session (see Study 1 for a description). To enable comparison of the observer and therapist ratings, the DBT ACS items that make up the briefer DBT AC-I were recoded to a binary scale to match the scale of the DBT AC-I. Specifically, DBT ACS ratings of 0, 4, and 5 were recoded to 1 (adherent) and ratings of 1, 2, and 3 were recoded to 0 (non-adherent).

Session Targets After each session, therapists completed a brief survey to assess their clients' target behaviors since the last session. The 11-item Borderline Symptom List—Behavioral Items (BSL-BI; Bohus et al., 2009) assessed a variety of impulsive behaviors common in BPD (e.g., self-injury, substance use, binge-eating) on a scale from 0 (not at all) to 4 (daily or more often). The 23-item version of the Therapist Interview (TI; Chalker et al., 2015) assessed three types of challenging client behaviors common in BPD that are based on the DBT concept of therapy-interfering behaviors: (1) interpersonal negativity (e.g., "behaved in an inflexible or defiant manner"), (2) avoidant/disengaged (e.g., "missed session without calling"), and (3) behavioral dysregulation (e.g., "arrived at sessions under the influence of drugs or alcohol"). All items were rated on a 5-point scale from 0 (behavior did not occur) to 4 (behavior occurred and seriously interfered). A single item also assessed therapists' overall level of stress and/

or burnout in treating the client from 1 (none at all) to 5 (extreme).

Data Analysis

Aim 1 evaluated the psychometric properties of the therapist-rated version of the DBT AC-I. A confirmatory factor analysis (CFA) was tested to determine if the DBT AC-I items conformed to a single dimension consistent with the scoring of the DBT ACS. To determine goodness of fit, the following indices and thresholds were used: Root Mean Square Error of Approximation (RMSEA < 0.06; Hu & Bentler, 1999), Comparative Fit Index (CFI > 0.95; Byrne, 2006), Normed Fit Index (NFI > 0.90; Bentler & Bonett, 1980), and Nonnormed Fit Index (NNFI > 0.90; Bentler & Bonett, 1980).

Therapist-observer concordance was evaluated at the item-level using the Agreement Corrected 1 (AC_1) statistic, which is recommended as an alternative to kappa when rater agreement and trait prevalence are both high (Gwet, 2008). The AC_1 statistic resolves the well-known paradox in which kappa appears small when rater agreement is high (Cicchetti & Feinstein, 1990). The AC_1 statistic is interpreted using the same criteria for interpreting kappas where less than 0.20 is poor, 0.21–0.40 is fair, 0.41–0.60 is moderate, 0.61–0.80 is good, and 0.81–1.00 is very good (Altman, 1999). We set a benchmark for retaining items as achieving at least moderate therapist-observer agreement ($AC_1 \geq 0.41$). To provide an index of overall therapist-observer concordance, an ICC was computed for the DBT AC-I total score using the random-effects estimate with two raters pooled (ICC [2, 2]; Shrout & Fleiss, 1979). According to Cicchetti's (1994) criteria for interpreting ICCs, less than 0.40 is poor, 0.40–0.59 is fair, 0.60–0.74 is good, and 0.75–1.00 is excellent.

Convergent validity was evaluated by examining the correlation between the DBT AC-I total score and the DBT ACS computed global score. Criterion validity was evaluated by determining how well the DBT AC-I identified adherent versus non-adherent sessions as defined by the gold standard DBT ACS (i.e., computed global score ≥ 4.0). All possible cut-off scores were examined in terms of sensitivity (i.e., the percent of true positives that are above the cut-off) and specificity (i.e., the percent of true negatives that are below the cut-off). As the cut-off score varies, the locus (1-specificity, sensitivity) yields a receiver operating characteristic (ROC) curve that displays the performance of all possible cut-off scores and the optimal score was selected by the max–min approach that simultaneously maximizes the probability of predicting a true-positive or a true-negative (Gallop et al., 2003). The AUC was calculated to quantify the ability of the DBT AC-I to correctly classify adherent versus non-adherent sessions and an AUC ≥ 0.90 indicates excellent criterion validity. Content validity was evaluated by examining whether the items included in the DBT AC-I adequately represented the strategies most often in need of improvement among DBT therapists in routine practice. Specifically, the 66 items in the DBT ACS were examined to determine if those with high rates of non-adherence (defined as ≥ 25% of sessions) were included in the DBT AC-I.

Aim 2 sought to identify therapist-, client-, and session-level predictors of therapist accuracy in rating their adherence to DBT. Therapist factors included demographics (sex, age, degree), DBT experience (workshop days, years providing DBT, self-efficacy), and DBT knowledge. Client factors included psychiatric diagnoses (BPD, total diagnoses), SIB history (C-SSRS), and functioning (GAS score). Session factors included session type (telehealth vs. in-person), target behaviors (BSL-BI, TI), and observer-rated adherence (DBT ACS). Therapist accuracy was calculated as the percentage of concordant items between therapist and observer ratings. One subject with an extreme outlying low value was capped at the value of the next closest subject through winsorization. Mixed effects models were used to accommodate the nested nature of the data (sessions within client-therapist dyads). Initial univariate mixed effects models were conducted to examine the effect of each predictor on therapist accuracy. A final multivariate mixed effects model was run that included all predictors found to be significant in the univariate models.

For Aim 3, the structural, convergent, and criterion validity of the observer-rated DBT AC-I were evaluated using the same analytic approaches as described in Aim 1. Interrater reliability (ICC) was evaluated between observers and the gold standard rater of the DBT ACS using a random subset of 10% of all coded sessions ($n = 10$).

Results

Preliminary Descriptive Analyses

Characteristics of therapists and clients are shown in Tables 1 and 2, respectively. Due to the COVID-19 pandemic, most sessions ($n = 81$, 81.0%) were conducted via teletherapy and the remainder ($n = 19$, 19.0%) occurred in person. On average, the sessions occurred after clients had been receiving DBT for 10.0 months ($SD = 8.7$, range = 1–40). Potential session targets reported on the BSL-BI included suicide attempts ($n = 1$, 1.0%), self-harm ($n = 6$, 6.0%), telling other people they were going to kill themselves ($n = 3$, 3.0%), binge-eating ($n = 17$, 17.0%), vomiting/purging ($n = 8$, 8.0%), getting drunk ($n = 20$, 20.0%), taking drugs ($n = 8$, 8.0%), misusing prescribed medications ($n = 5$, 5.0%), other high-risk behaviors ($n = 2$, 2.0%), anger outbursts or physically attacking others ($n = 14$, 14.0%), and impulsive sexual encounters ($n = 2$, 2.0%). The most common therapy-interfering behaviors reported on the TI were not

Table 2 Study 2 client characteristics at baseline

	Clients (*n* = 50)
Age, *M* ± *SD*	32.9 ± 11.2
Gender	
Female	42 (84.0)
Male	6 (12.0)
Transgender	1 (2.0)
Non-binary	1 (2.0)
Race/Ethnicity	
Non-Hispanic White	41 (82.0)
Hispanic White	3 (6.0)
Asian/Asian-American	3 (6.0)
Black/African-American	2 (4.0)
Multi-racial	1 (2.0)
Self-Injurious Thoughts and Behaviors (past 3 mos)	
Active suicidal thoughts	27 (54.0)
Non-suicidal self-injury	17 (34.0)
Suicide attempt[a]	5 (10.0)
Current Psychiatric Diagnoses	
Borderline personality disorder	27 (54.0)
Posttraumatic stress disorder	27 (54.0)
Major depressive disorder	27 (54.0)
Bipolar I or II disorder	13 (26.0)
Any anxiety disorder	23 (46.0)
Any substance use disorder	15 (30.0)
Any eating disorder	11 (22.0)
GAS score, *M* ± *SD*	54.5 ± 12.3
Current DBT Modes	
DBT individual therapy	50 (100.0)
DBT skills group	36 (72.0)
DBT phone coaching	44 (88.0)

Data are presented as *n* (%) unless otherwise noted

GAS Global Assessment Scale

[a] Includes actual, aborted, and interrupted suicide attempts

Table 3 Descriptive data for DBT AC-I therapist and observer ratings

	Therapist Rating *M (SD)*	Observer Rating *M (SD)*
Diary card**	0.90 (0.30)	0.92 (0.27)
Organize by targets*	0.93 (0.26)	0.81 (0.39)
Emotion focus*	0.93 (0.26)	0.84 (0.37)
Describe specifically*	0.88 (0.33)	0.84 (0.37)
Chain analysis	0.72 (0.45)	0.79 (0.41)
Teach new information	0.97 (0.17)	0.96 (0.20)
Generate solutions*	0.88 (0.33)	0.81 (0.39)
Activate new behavior*	0.81 (0.39)	0.64 (0.48)
Provide coaching feedback	0.88 (0.33)	0.90 (0.30)
Generalize new learning**	0.87 (0.34)	0.72 (0.45)
Reinforcement*	0.96 (0.20)	0.99 (0.10)
Aversive contingencies	0.89 (0.31)	0.97 (0.17)
Informal exposure	0.90 (0.30)	0.95 (0.22)
Challenge cognitions	0.90 (0.30)	0.97 (0.17)
Validation level 4	0.92 (0.27)	0.99 (0.10)
Validation level 5*	0.97 (0.17)	0.90 (0.30)
Validation level 6*	0.99 (0.10)	0.92 (0.27)
Warm engagement*	0.99 (0.10)	0.99 (0.10)
Self-disclosure	0.91 (0.29)	0.99 (0.10)
Direct confrontation	0.89 (0.31)	0.87 (0.34)
Unorthodox irreverence	0.97 (0.17)	1.00 (0.00)
Balanced style and strategies*	0.86 (0.35)	0.83 (0.38)
Model dialectical thinking*	0.84 (0.37)	0.65 (0.48)
Consultation to the client	0.98 (0.14)	1.00 (0.00)
Suicidal behaviors protocol	0.93 (0.26)	0.96 (0.20)
Total Score	22.67 (2.37)	22.21 (2.72)

Mean scores reflect the proportion of sessions in which the strategy was rated as adherent

*Indicates a required strategy in individual DBT

**Indicates a required strategy in individual DBT after pre-treatment

completing homework assignments (37%), failing to make eye contact (35%), and withdrawing or behaving in an inattentive or apathetic manner (23%). On average, therapists reported feeling 'not at all' to 'slightly' stressed/burned out in treating the client ($M = 1.50$, $SD = 0.72$). Across sessions, the average observer-rated DBT ACS computed global score was 3.97 ($SD = 0.18$, range = 3.28–4.31) indicating that therapists delivered DBT slightly below adherence on average. Overall, 56% of the sessions were adherent and 44% were non-adherent according to the DBT ACS.

Aim 1. Psychometric Properties of the DBT AC-I as a Therapist Self-Report Measure

Therapists completed the DBT AC-I immediately after the session ($n = 31$, 31.0%), the same day as the session but not immediately after ($n = 48$, 48.0%), and one or more days after the session ($n = 21$, 21.0%). The median time to complete the DBT AC-I was 20.0 min. Prior to completing the checklist, most therapists read the DBT AC-I manual completely ($n = 43$, 86.0%) and the remainder read the manual partially ($n = 6$, 12.0%) or not at all (n = 1, 2.0%). Therapists rated the manual as 'very helpful' on average ($M = 3.9$ out of 4, $SD = 1.0$) and the DBT AC-I as having above average usability on the SUS ($M = 71.8$, $SD = 10.8$).

Structural Validity

A CFA with one dimension that included all the items yielded an acceptable fit: RMSEA = 0.059, CFI = 0.958, NFI = 0.935, NNFI = 0.821. This indicates that the use of a total summed score adequately represents the data.

Table 4 Concordance between therapist and observer ratings on the DBT AC-I items

	Frequencies ($n = 100$ sessions)				Concordance	
	True Positive	True Negative	False Positive	False Negative	% Agreement	AC_1
Diary card	85	3	5	7	88.0	0.86
Organize by targets	75	1	18	6	76.0	0.69
Emotion focus	77	0	16	7	77.0	0.71
Describe specifically	74	2	14	10	76.0	0.68
Chain analysis	57	6	15	22	63.0	0.41
Teach new information	93	0	4	3	93.0	0.92
Generate solutions	77	8	11	4	85.0	0.80
Activate new behavior	56	11	25	8	67.0	0.45
Provide coaching feedback	81	3	7	9	84.0	0.80
Generalize new learning	64	5	23	8	69.0	0.54
Reinforcement	95	0	1	4	95.0	0.95
Aversive contingencies	88	2	1	9	90.0	0.88
Informal exposure	85	0	5	10	85.0	0.83
Challenge cognitions	87	0	3	10	87.0	0.85
Validation level 4	91	0	1	8	91.0	0.90
Validation level 5	89	2	8	1	91.0	0.90
Validation level 6	91	0	8	1	91.0	0.90
Warm engagement	98	0	1	1	98.0	0.98
Self-disclosure	90	0	1	9	90.0	0.89
Direct confrontation	78	2	11	9	80.0	0.75
Unorthodox irreverence	97	0	0	3	97.0	0.97
Balanced style and strategies	73	4	13	10	77.0	0.69
Model dialectical thinking	58	9	26	7	67.0	0.47
Consultation to the client	98	0	0	2	98.0	0.98
Suicidal behaviors protocol	90	1	3	6	91.0	0.90

"True positive" = rated as adherent by both therapists and observers

"True negative" = rated as non-adherent by both therapists and observers

"False positive" = rated as adherent by therapists and non-adherent by observers

"False negative" = rated as non-adherent by therapists and adherent by observers

Therapist-Observer Concordance

Descriptive data for the therapist and observer ratings for each DBT AC-I item and the total score are shown in Table 3. Table 4 reports the frequencies of concordant and discordant ratings, agreement rates, and AC_1 statistics for each DBT AC-I item. The average agreement rate across all items was 84.2% (range = 63.0–98.0%, $SD = 0.10$). All 25 items had at least moderate therapist-observer concordance ($AC_1 \geq 0.41$), which we considered the minimum acceptable threshold of agreement. Concordance was very good for 14 items (56%), good for 7 items (28%), and moderate for 4 items (16%). Therapists accurately rated themselves as adherent on 81.9% of items and non-adherent on 2.4% of items. Therapists overestimated their adherence on 8.8% of items and underestimated their adherence on 7.0% of items. Given that therapists over- and under-estimated their adherence at similar rates, these discordant ratings canceled themselves out such that the average DBT AC-I total score did not significantly differ between therapists and observers (M difference = 0.46, range = − 8–14, $SD = 3.52$, t (99) = 1.31, $p = 0.19$, $d = 0.13$). However, therapist-observer concordance for the DBT AC-I total score was poor (ICC = 0.09, 95% CI − 0.35–0.39) due to the variability in agreement rates across therapists.

Convergent Validity

The therapist-rated DBT AC-I total score was not significantly correlated with the observer-rated DBT ACS computed global score ($r = 0.05$, $p = 0.64$), indicating it had poor convergent validity with the gold standard measure of adherence to DBT.

Criterion Validity

The AUC for the DBT AC-I total score was 0.54 (OR = 1.05, 95% CI 0.89–1.24) and non-significant (χ^2 (1) = 0.30, $p = 0.58$), indicating it had poor criterion validity for predicting adherent versus non-adherent sessions according to the observer-rated DBT ACS. Classification accuracy was only slightly better than chance (52.0%; 95% CI 41.8–62.1%; sensitivity = 56.8%; specificity = 48.2%).

Content Validity

Four DBT ACS items met the threshold for having a high rate of non-adherence (≥ 25% of sessions), of which 3 were included in the 25-item DBT AC-I (generalizes learning, dialectical synthesis, activates new behavior) and 1 was not (troubleshooting). In DBT, troubleshooting is required whenever a commitment is obtained. Overall, the contingent strategies of asking for and troubleshooting a commitment were below adherence in 75.0% of the sessions, suggesting that adding these strategies to the DBT AC-I would improve its content validity as a measure designed to help therapists identify and improve areas of non-adherence.

Aim 2. Predictors of Therapist Accuracy of Self-Rated Adherence to DBT

Therapist accuracy (i.e., the percentage of concordant items between therapists and observers) ranged from 68 to 100% across therapists ($M = 84.4\%$, $SD = 0.08$). Table 5 reports the results of univariate mixed effects models predicting therapist accuracy.

Table 5 Client, therapist, and session predictors of therapist accuracy in self-rating adherence

Predictors	B	SE	η^2
Therapist Factors			
Sex (ref = female)	− 0.023	0.032	0.007
Age	− 0.002	0.001	0.051
Degree (ref = doctorate)	− 0.038	0.026	0.030
# days of DBT workshops	− 0.001	0.001	0.009
# years of DBT experience	− 0.004	0.003	0.022
Self-efficacy in DBT	− 0.026	0.026	0.014
DBT knowledge test	0.232*	0.103	0.065
Client Factors			
BPD diagnosis	− 0.018	0.023	0.009
# of psychiatric diagnoses	0.004	0.007	0.004
C-SSRS any SIB (past 3 mos)	− 0.018	0.024	0.008
C-SSRS suicidal ideation	0.017*	0.007	0.076
GAS score	− 0.001	0.001	0.034
Session Factors			
Session type (ref = telehealth)	− 0.046	0.029	0.035
BSL-BI	− 0.034	0.050	0.005
TI	0.043	0.059	0.008
DBT ACS computed global score	0.296***	0.048	0.290

Each row is from a separate mixed effects model

DBT dialectical behavior therapy, *BPD* borderline personality disorder, *C-SSRS* Columbia Suicide Severity Rating Scale, *GAS* Global Assessment Scale, *SIB* self-injurious behavior, *BSL-BI* Borderline Symptom List-Behavioral Items, *TI* Therapist interview, *DBT ACS* Dialectical Behavior Therapy Adherence Coding Scale

*$p < .05$, **$p < .01$, ***$p < .001$

Therapist Factors

Therapists with greater knowledge of DBT were significantly more accurate in rating their own adherence ($p = 0.03$). Therapist sex, age, degree, number of DBT workshop days, years of DBT experience, and self-efficacy did not significantly predict accuracy.

Client Factors

Therapists were significantly more accurate when rating sessions conducted with clients with more severe suicidal ideation ($p = 0.02$). Client diagnoses, SIB history, and functioning did not significantly predict therapist accuracy.

Session Factors

Therapists with higher observer-rated adherence on the DBT ACS were significantly more accurate in rating their own adherence ($p < 0.0001$). Session type and target behaviors did not significantly predict therapist accuracy.

Combined Model

A multivariate MEM that included the three significant predictors from the individual models was significant (Likelihood χ^2 (2) = 8.01, $p = 0.019$). In this model, only greater observer-rated adherence remained a significant predictor of higher therapist accuracy (B = 0.277, SE = 0.050, $p < 0.0001$). Therapist accuracy was not significantly related to knowledge of DBT ($p = 0.76$) or the severity of the client's suicidal ideation ($p = 0.11$).

Aim 3. Psychometric Properties of the DBT AC-I as an Observer-Rated Measure

All analyses utilized the observer ratings on the DBT ACS items that make up the briefer DBT AC-I. Each DBT ACS item was recoded to a binary scale to match the scale of the DBT AC-I. We examined the psychometric properties of two potential versions of the observer-rated measure: (1) a 25-item version that was identical to the DBT AC-I completed by therapists, and (2) a 26-item version that added the commitment/troubleshooting item identified as important to improve the content validity of the measure (see Aim 1 results above).

Interrater Reliability

The average item-level agreement rate between observers and the gold standard rater was 94.8% for the 25-item version and 95.0% for the 26-item version. Inter-rater reliability was excellent for the DBT AC-I total score for both the 25-item version (ICC = 0.93, 95% CI 0.71–0.98) and the 26-item version (ICC = 0.91, 95% CI 0.64–0.98).

Structural Validity

CFAs with one dimension that included all the items indicated adequate fit for both the 25-item version (RMSEA = 0.06, CFI = 0.96 NFI = 0.90, NNFI = 0.93) and the 26-item version (RMSEA = 0.09, CFI = 0.95, NFI = 0.86, NNFI = 0.91). Although the RMSEA and NFI fit indexes were slightly above/below the recommended thresholds for one or both versions, this is quite common in smaller samples ($n \leq 100$; Taasoobshirazi & Wang, 2016). Thus, the use of a total summed score appears to adequately represent the data.

Convergent Validity

The DBT AC-I total score was significantly correlated with the DBT ACS computed global score for the 25-item ($r = 0.89$, $p < 0.001$) and the 26-item version ($r = 0.90$, $p < 0.001$).

Criterion Validity

Both versions of the observer-rated DBT AC-I were used to predict adherent versus non-adherent sessions according to the DBT ACS computed global score. For the 25-item version, the AUC was 0.93 and maximum accuracy was achieved at a cut-off score of 23 (85% overall agreement; 95% CI 76.5–91.4%; sensitivity = 89.3%; specificity = 79.6%). For the 26-item version, the AUC was 0.94 and the optimal cut-off score was 23 (86% overall agreement; 95% CI 77.6–92.1%; sensitivity = 91.1%, specificity = 79.6%). This indicates that the 26-item version increases the ability to identify adherent sessions to an excellent level (sensitivity ≥ 0.90), whereas the 25-item version is slightly below this threshold.

Discussion

These studies aimed to develop and test a pragmatic measure of adherence to DBT individual therapy that would enable therapists, programs, and systems to evaluate and improve the quality of services they deliver. In Study 1, the DBT AC-I was empirically derived from the gold standard DBT ACS using a large archival database to identify critical items and an optimal rating scale. To maximize construct validity, content for the items and an accompanying manual was written to be consistent with the definitions used in the DBT ACS. The measure and manual were then iteratively revised based on feedback from DBT therapists and experts until high usability and understandability were achieved. In Study 2, the psychometrics of the DBT AC-I as a therapist self-report and an observer-rated measure were evaluated among 50 therapist-client dyads engaged in DBT in routine practice settings. Results have important implications for the implementation and practice of DBT.

When used by therapists to rate their own adherence, the DBT AC-I was efficient (median time to complete was 20 min) and rated as highly usable. Concordance between therapist and observer ratings was at least moderate for all items and very good for more than half the items. Agreement tended to be lower for items that were more frequently non-adherent (e.g., chain analysis, activating new behavior, modeling dialectical thinking) and higher for items with the highest rates of adherence (e.g., warm engagement, validation, reinforcement). This suggests that therapists may be more accurate in evaluating their own adherence for strategies that they are more proficient in delivering. Alternatively, given prior research indicating therapists tend to overestimate their adherence (e.g., Brookman-Frazee et al., 2021; Hogue et al., 2015; Hurlburt et al., 2010), agreement may be higher when items are adherent due to therapist rating bias. Notably, therapists in the present study overestimated and underestimated their adherence at similar rates. As a result, the average DBT AC-I total scores generated by therapists and observers did not significantly differ even though there was poor overall therapist-observer concordance. Additionally, the therapist-rated DBT AC-I total score had poor convergent validity with the gold standard DBT ACS and was unable to distinguish between adherent and non-adherent sessions. These findings are consistent with the larger literature indicating that therapist self-report is an efficient method of evaluating adherence but often has limited effectiveness due to therapists' difficulty accurately self-rating their adherence (e.g., Brosan et al., 2008; Hogue et al., 2015; Hurlburt et al., 2010; Martino et al., 2009).

Importantly, therapist-observer concordance varied considerably across therapists with overall accuracy rates ranging from 68 to 100%. Higher therapist accuracy was predicted by greater DBT knowledge and observer-rated adherence. More knowledgeable and adherent therapists may be more able to deliver the strategies they intended to carry out and self-reflect on their performance, both of which predict greater accuracy in self-reported adherence in other EBPs (Brookman-Frazee et al., 2021). Therapists were also more accurate when rating sessions with clients who had more severe suicidal ideation. DBT sessions often have a clearer structure when clients' suicide risk increases (e.g., certain strategies and protocols are required), and therapists have been found to rate their adherence more accurately when delivering EBPs with more structured content (Brookman-Frazee et al., 2021). Overall, these findings indicate that the DBT AC-I may be both efficient and effective as a self-report measure of adherence for some but not all therapists. Given that there is no way to easily predict which therapists are likely to be accurate raters, DBT AC-I scores generated by therapist self-report should not be assumed to be reliable. Future research may be able to develop algorithms to improve the accuracy of therapists' self-rated adherence on the DBT AC-I by adjusting scores based on key predictors.

When rated by trained observers, the DBT AC-I had excellent interrater reliability as well as strong convergent and criterion validity with the gold standard DBT ACS. Of the two versions that were evaluated, the 26-item DBT AC-I is recommended given its improved content validity and excellent ability to identify adherent sessions (sensitivity) compared to the 25-item version. Notably, these strong psychometric properties were achieved despite eliminating more than 60% of the DBT ACS items (from 66 to 26) and simplifying the rating scale (from a 6-point to a binary scale). In addition, the DBT AC-I is free and publicly available (whereas the DBT ACS is not), further reducing burden for users. Altogether, the DBT AC-I offers a more efficient and comparably effective alternative to the gold standard DBT ACS when used by trained observers. Nonetheless, any method that relies on direct observation of sessions by trained raters is likely to be infeasible in some settings (e.g., agencies that do not have the equipment needed to record sessions or access to trained observers).

Limitations

Results must be interpreted in the context of several limitations. First, the participants in these studies may not be representative of DBT therapists and clients in routine practice settings more broadly. To maximize generalizability, we recruited therapists with a range of degrees, disciplines, training backgrounds, and practice settings and did not use inclusion/exclusion criteria for clients except to ensure they were not minors. However, therapists who enrolled in the study were typically highly trained, experienced in DBT, and worked primarily in private practice settings. In addition, therapists may have been particularly motivated to learn about and improve their adherence to DBT. Clients were moderately impaired on average, had relatively low rates of recent self-injurious behavior, and consented to have their sessions reviewed by the study team, which may not be typical of DBT clients more broadly. It is also unknown if the results would vary if the DBT AC-I were used to rate sessions with adolescent clients, although prior research has found comparable reliability and validity of the DBT ACS in adult and adolescent samples (Harned et al., 2021a, b). Additional research with larger and more diverse therapist and client samples is needed to replicate and extend the present findings.

Other limitations include that most sessions were conducted via telehealth due to the COVID-19 pandemic and it is possible that this impacted results; however, session type did not predict therapist accuracy in self-rated adherence. Additionally, to enable evaluation of convergent and criterion validity with the gold standard measure, the trained observers completed the full DBT ACS rather than just the items included in the DBT AC-I. It is possible that ratings may have differed if only the DBT AC-I items were completed. The use of a binary rating scale (adherent vs. non-adherent) for the DBT AC-I also limits the ability of the measure to evaluate therapist competence compared to the continuous scale of the DBT ACS. Finally, this study focused only on developing a pragmatic measure of adherence to DBT individual therapy. Future research is needed to develop measures of adherence for the other modes of DBT (e.g., group therapy).

Implications and Next Steps

The lack of a pragmatic adherence measure has been a significant barrier to DBT implementation and quality assurance efforts. Without such a measure, there has been no way to evaluate if DBT is being delivered as intended to the clients who need it. This is particularly concerning given that therapists in routine practice often deliver DBT non-adherently according to this study (44% of sessions) and prior research (48% of sessions; Harned et al., 2021a, b). Moreover, many clients who receive DBT are at high risk for suicide and therapist adherence is an important mechanism by which suicidal behavior is reduced (Harned et al., 2022). The DBT AC-I represents a significant advance towards enabling stakeholders to evaluate if the treatment being provided is consistent with DBT as it is defined in the treatment manual. The final 26-item DBT AC-I has several notable strengths, including being brief, easy to use, freely available (at www.dbtadherence.com), and having excellent construct and content validity. However, other indices of effectiveness (interrater reliability, convergent and criterion validity) vary depending on the characteristics of the rater, indicating a need for caution in how the measure is used.

Given that therapists overall had difficulty generating reliable and valid scores on the DBT AC-I, therapists' self-rated adherence on the DBT AC-I should not be assumed to reflect their actual adherence to DBT. These findings suggest that the brief training provided to therapists in this study (i.e., reading the DBT AC-I manual) is likely to be insufficient for many therapists to complete the DBT AC-I accurately. Consequently, we have since created a set of expert-rated mock sessions for therapists to use for training and calibration purposes (available at www.dbtadherence.com). We also made minor revisions to the measure and manual prior to making it publicly available with the goal of increasing the reliability of therapist ratings (e.g., clarifying definitions and giving more examples for items with lower agreement rates). An important next step for research will be to evaluate whether these revisions and training methods improve the effectiveness of the DBT AC-I as a therapist self-report measure. Until then, it is recommended that the DBT AC-I be used for lower stakes purposes (e.g., quality improvement, supervision and training), but not as a formal assessment of adherence, when completed by therapists (and others) whose reliability has not been established.

In contrast, when completed by trained observers the DBT AC-I had excellent psychometric properties but was more burdensome than therapist self-report, reflecting the tension between effectiveness and efficiency that is often encountered in adherence measurement (Schoenwald et al., 2011). Future research is needed to identify methods for assessing adherence to DBT that are effective and less resource-intensive than direct observation of sessions. For example, recent studies have found that behavioral rehearsal (Becker-Haimes et al., 2022) and review of clinical worksheets (Wiltsey Stirman et al., 2021) hold promise as indirect methods of evaluating therapist adherence. Thus, while the DBT AC-I represents an important advance in DBT practice and research, additional work to synthesize the tension between efficiency and effectiveness is needed.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10488-023-01274-x.

Funding This research was supported by a grant from the National Institute of Mental Health awarded to the first author [R56MH117085]. We would like to thank the Principal Investigators who allowed archival data from their clinical trials to be used in the initial scale development analyses, including Drs. Marsha Linehan, Elizabeth McCauley, Michelle Berk, Joan Asarnow, Stefan Priebe, and Thomas Lynch. The clinical trials from which the archival data were derived were supported by grants from the National Institute of Mental Health [R01MH034486, PI: Dr. Linehan; R01MH090159, PIs: Drs. Linehan and McCauley; R01MH93898, Drs. Berk and Asarnow; R34MH082143, PI: Dr. Harned; R34MH106598, PI: Dr. Harned], the National Institute on Drug Abuse [R01DA014997, PIs: Drs. Linehan and Lynch], and the U.K. National Institute for Health Research [PB-PG-0906-10540, PI: Dr. Priebe]. We would also like to thank Dr. Shannon Dorsey, Isabelle Tully, Sophia Robinson, Jesse Homan, and the research participants for their contributions to this project. Drs. Harned, Schmidt, and Korslund are paid to provide training and consultation in DBT.

References

Altman, D. G. (1999). *Practical statistics for medical research*. Hall/CRC Press.

An, X., & Yung, Y. F. (2014). *Item response theory: What it is and how you can use the IRT procedure to apply it*. SAS Institute.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction, 24*, 574–594. https://doi.org/10.1002/jclp.20221

Becker-Haimes, E. M., Marcus, S. C., Klein, M. R., Schoenwald, S. K., Fugo, P. B., & Beidas, R. S. (2022). A randomized trial to identify accurate measurement methods for adherence to cognitive-behavioral therapy. *Behavior Therapy*. https://doi.org/10.1016/j.beth.2022.06.001

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–600. https://doi.org/10.1037/0033-2909.88.3.588

Bohus, M., Kleindienst, N., Limberger, M. F., Stieglitz, R., & DomsallaWolf, M. M. (2009). The short version of the borderline symptom list (BSL-23): Development and initial data on psychometric properties. *Psychopathology, 42*, 32–39. https://doi.org/10.1159/000173701

Brookman-Frazee, L., Stadnick, N. A., Lind, T., Roesch, S., Terrones, L., & Lau, A. S. (2021). Therapist-observer concordance in ratings of EBP strategy delivery: Challenges and targeted directions in pursuing pragmatic measurement in children's mental health services. *Administration and Policy in Mental Health Services Research, 48*, 155–170. https://doi.org/10.1007/s10488-020-01054-x

Brosan, L., Reynolds, S., & Moore, R. G. (2008). Self-evaluation of cognitive therapy performance: Do therapists know how competent they are? *Behavioural and Cognitive Psychotherapy, 36*, 581–587. https://doi.org/10.1017/S1352465808004438

Carmel, A., Rose, M., & Fruzzetti, A. (2014). Barriers and solutions to implementing dialectical behavior therapy in a public behavioral health system. *Administration and Policy in Mental Health Services Research, 41*, 608–614. https://doi.org/10.1007/s10488-013-0504-6

Chalker, S. A., Carmel, A., Atkins, D. C., Landes, S. J., Kerbrat, A. H., & Comtois, K. A. (2015). Examining challenging behaviors of clients with borderline personality disorder. *Behaviour Research and Therapy, 75*, 11–19. https://doi.org/10.1016/j.brat.2015.10.003

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low Kappa: II. Resolving the paradoxes. *Journal of*

Clinical Epidemiology, 43, 551–558. https://doi.org/10.1016/0895-4356(90)90159-m

DeCou, C. R., Comtois, K. A., & Landes, S. J. (2019). Dialectical behavior therapy is effective for the treatment of suicidal behavior: A meta-analysis. *Behavior Therapy, 50*, 60–72. https://doi.org/10.1016/j.beth.2018.03.009

DuBose, A. P., Botanov, Y., & Ivanoff, A. (2019). International implementation of dialectical behaviour therapy: The challenge of training therapists across cultures. In M. A. Swales (Ed.), *The Oxford handbook of dialectical behaviour therapy* (pp. 909–930). Oxford University Press.

Endicott, J., Spitzer, R. L., Fleiss, J. L., & Cohen, J. (1976). The global assessment scale. A procedure for measuring overall severity of psychiatric disturbance. *Archives of General Psychiatry, 33*, 766–771. https://doi.org/10.1001/archpsyc.1976.01770060086012

Flynn, D., Joyce, M., Gillespie, C., Kells, M., Swales, M., & Weihrauch, M. (2020). Evaluating the national multisite implementation of dialectical behaviour therapy in a community setting: A mixed methods approach. *BMC Psychiatry, 20*, 235. https://doi.org/10.1186/s12888-020-02610-3

Gallop, R. J., Crits-Christoph, P., Muenz, L. R., & Tu, X. M. (2003). Determination and interpretation of the optimal operating point for ROC curves derived through generalized linear models. *Understanding Statistics, 2*, 219–242. https://doi.org/10.1207/S15328031US0204_01

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *The British Journal of Mathematical and Statistical Psychology, 61*, 29–48. https://doi.org/10.1348/000711006X126600

Harned, M. S., Gallop, R. J., Schmidt, S. C., & Korslund, K. E. (2022). The temporal relationships between therapist adherence and client outcomes in dialectical behavior therapy. *Journal of Consulting and Clinical Psychology, 90*, 272–281. https://doi.org/10.1037/ccp0000714

Harned, M. S., Korslund, K. E., Schmidt, S. C., & Gallop, R. J. (2021a). The Dialectical Behavior Therapy Adherence Coding Scale (DBT ACS): Psychometric properties. *Psychological Assessment, 33*, 552–561. https://doi.org/10.1037/pas0000999

Harned, M. S., Schmidt, S. C., & Korslund, K. E. (2021b). Does adding the dialectical behavior therapy prolonged exposure (DBT PE) protocol for PTSD to DBT improve outcomes in public mental health agencies? A pilot nonrandomized effectiveness trial with benchmarking. *Behavior Therapy, 52*, 639–655. https://doi.org/10.1016/j.beth.2020.08.003

Herschell, A. D., Lindhiem, O. J., Kogan, J. N., Celedonia, K. L., & Stein, B. D. (2014). Evaluation of an implementation initiative for embedding dialectical behavior therapy in community settings. *Evaluation and Program Planning, 43*, 55–63. https://doi.org/10.1016/j.evalprogplan.2013.10.007

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1*, 104–121. https://doi.org/10.1177/109442819800100106

Hogue, A., Dauber, S., Lichvar, E., Bobek, M., & Henderson, C. E. (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 Services Research, 42*, 229–243. https://doi.org/10.1007/s10488-014-0548-2

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. https://doi.org/10.1080/10705519909540118

Hurlburt, M. S., Garland, A. F., Nguyen, K., & Brookman-Frazee, L. (2010). Child and family therapy process: Concordance of therapist and observational perspectives. *Administration and Policy in Mental Health Services Research, 37*, 230–244. https://doi.org/10.1007/s10488-009-0251-x

Linehan, M. M. (1993). *Cognitive-behavioral treatment of borderline personality disorder*. Guilford Press.

Linehan, M. M., & Korslund, K. E. (2003). *Dialectical behavior therapy adherence coding scale manual*. University of Washington.

Loades, M. E., & Myles, P. J. (2016). Does a therapist's reflective ability predict the accuracy of their self-evaluation of competence in cognitive behavioural therapy? *The Cognitive Behaviour Therapist, 9*, 1–14. https://doi.org/10.1017/S1754470X16000027

Martino, S., Ball, S., Nich, C., Frankforter, T. L., & Carroll, K. M. (2009). Correspondence of motivational enhancement treatment integrity ratings among therapists, supervisors, and observers. *Psychotherapy Research, 19*, 181–193. https://doi.org/10.1080/10503300802688460

Navarro-Haro, M. V., Harned, M. S., Korslund, K. E., DuBose, A., Chen, T., Ivanoff, A., & Linehan, M. M. (2019). Predictors of adoption and reach following dialectical behavior therapy intensive training. *Community Mental Health Journal, 55*, 100–111. https://doi.org/10.1007/s10597-018-0254-8

Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology, 75*, 829–841. https://doi.org/10.1037/0022-006X.75.6.829

Posner, K., Brown, G., StanleyBrentYershova, K. D. A. K. V., & Mann, J. J. (2011). The Columbia-Suicide Severity Rating Scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American Journal of Psychiatry, 168*, 1266–1277. https://doi.org/10.1176/appi.ajp.2011.10111704

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*, 24–34. https://doi.org/10.1007/s10488-008-0197-4

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553

Schoenwald, S. K., & Garland, A. F. (2013). A review of treatment adherence measurement methods. *Psychological Assessment, 25*, 146–156. https://doi.org/10.1037/a0029715

Schoenwald, S. K., Garland, A. F., Chapman, J. E., Frazier, S. L., Sheidow, A. J., & Southam-Gerow, M. A. (2011). Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health Services Research, 38*, 32–43. https://doi.org/10.1007/s10488-010-0321-0

Storebø, O. J., Stoffers-Winterling, J. M., Völlm, B. A., Kongerslev, M. T., Mattivi, J. T., Jørgensen, M. S., & Simonsen, E. (2020). Psychological therapies for people with borderline personality disorder. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD012955.pub2

Swales, M. A., Taylor, B., & Hibbs, R. A. B. (2012). Implementing dialectical behaviour therapy: Programme survival in routine healthcare settings. *Journal of Mental Health, 21*, 548–555. https://doi.org/10.3109/09638237.2012.689435

Taasoobshirazi, G., & Wang, S. (2016). The performance of the SRMR, RMSEA, CFI, and TLI: An examination of sample size, path size, and degrees of freedom. *Journal of Applied Quantitative Methods, 11*, 31–39.

Wiltsey Stirman, S., Gutner, C. A., Gamarra, J., Suvak, M. K., Vogt, D., & Resick, P. A. (2021). A novel approach to the assessment of fidelity to a cognitive behavioral therapy for PTSD using clinical worksheets: A proof of concept with cognitive processing therapy.

Behavior Therapy, 52, 656–672. https://doi.org/10.1016/j.beth.2020.08.005

Wolcott, M. D., & Lobczowski, N. G. (2021). Using cognitive interviews and think-aloud protocols to understand thought processes. *Currents in Pharmacy Teaching and Learning, 13*, 181–188. https://doi.org/10.1016/j.cptl.2020.09.005

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.