Administration and Policy in Mental Health and Mental Health Services Research
https://doi.org/10.1007/s10488-022-01219-w

ORIGINAL ARTICLE

Initial Development of Pragmatic Behavioral Activation Fidelity Assessments

Mary Beth Connolly Gibbons[1] · Jena Fisher[2] · Robert Gallop[3] · Eirini Zoupou[1] · Lang Duong[1] · Paul Crits-Christoph[1]

Accepted: 20 August 2022

Abstract

Purpose Our goal was to develop brief pragmatic assessments of Behavioral Activation (BA) fidelity to support its dissemination in low-resource settings.

Methods We used qualitative and quantitative methods across three investigations to develop pragmatic assessments rated from the perspective of therapists, patients, and observers: (1) we developed an initial comprehensive pool of 119 items and adapted/refined the item pool to 32 items through stakeholder focus groups and cognitive interviews; (2) independent blind judges rated each of items in the refined item pool on an early session of BA for 64 patients to support the selection of items based on predictive validity; and (3) we conducted a preliminary evaluation of the acceptability and feasibility of the assessments of BA fidelity from the perspective of therapists and patients.

Results The internal consistency reliability for the 10-item total score was .83 rated from the perspective of independent observers. The assessment was completed by patients following 90% of sessions and by clinicians following 93% of sessions. Items were rated high on overall satisfaction by both therapists ($M = 4.6$, $SD = 0.89$) and patients ($M = 4.8$, $SD = 0.41$).

Conclusion Our findings suggest that these brief assessments of BA fidelity are reliable, feasible, and acceptable to community stakeholders.

Keywords Behavioral activation · Fidelity · Community mental health · Pragmatic assessment · Evidence-based

Initial Development of Pragmatic Behavioral Activation Fidelity Assessments

Considerable effort and large sums of money have been spent over the last decade to disseminate evidence-based psychotherapies (EBPs) to usual care settings. Progress disseminating EBPs has been slow and our lack of reliable, valid, and scalable treatment fidelity measures for use in these settings impedes our ability to ensure patients are receiving effective interventions (Aarons et al., 2011; Schoenwald, 2011; Schoenwald et al., 2011; Southam-Gerow & McLeod, 2013). Numerous investigations have successfully used broad measures of fidelity characterizing general therapeutic strategies (Bearsley-Smith et al., 2008; Hepner et al., 2010; Kelley et al., 2010; Weersing et al., 2002) while others have focused on the development of measures to support the dissemination of specific EBPs (Chapman et al., 2013; Hogue et al., 2014). Because there has been a concerted effort to improve the quality of mental health services by disseminating specific EBPs to community settings (Creed et al., 2014; Karlin et al., 2010), further development of specific EBP fidelity scales that are scalable in these settings is essential.

There is a particular need for dissemination of EBPs to the publicly funded community mental health centers (CMHCs) in the United States, because millions of people utilize these settings for mental health services (Wells et al., 2010). In previous studies in the CMHC setting, we found response rates for treatment-as-usual (TAU) in the treatment of major depressive disorder (MDD) to be only 29% in one

✉ Mary Beth Connolly Gibbons
gibbonsm@pennmedicine.upenn.edu

[1] Department of Psychiatry, University of Pennsylvania, Philadelphia, PA, USA

[2] Merakey, Sharon Hill and Lafayette Hill, Lafayette Hill, PA, USA

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

Published online: 01 November 2022

investigation (Gibbons et al., 2012) and 13% in another (Connolly Gibbons et al., 2015). Efficacy trials suggest that outcomes in these settings can be improved through the provision of specific EBPs, yet CMHC's often lack the resources to train clinicians using the time-consuming and costly methods employed in efficacy trials. The challenge for CMHCs is how to train therapists to deliver evidence-based interventions with fidelity while minimizing cost.

Among EBPs, Behavioral Activation (BA) is a highly supported intervention for MDD that has the potential to improve outcomes in the CMHC setting. The essence of BA is a focus on increasing the patient's contact with sources of reward by helping them become more active and, consequently, improve one's life context. BA has demonstrated superiority to control treatments in multiple meta-analyses (Ekers et al., 2014; Stein et al., 2021) for patients with a primary diagnosis of MDD. Because of the relatively straightforward rationale for BA, this EBP is likely to be an appropriate psychotherapy for broad dissemination to master's level therapists typically employed by CMHCs. BA has also demonstrated effectiveness in the treatment of depression when delivered by non-specialists in primary care settings (Ekers et al., 2011) and as part of a home-based intervention delivered by social workers (Gitlin et al., 2013).

Many large-scale efforts to disseminate EBPs to community settings have used experts to ensure treatment integrity, including expert-led training workshops, expert consultation, and expert observer assessments of treatment fidelity. However, expert observer ratings are time-consuming, costly, and impractical for use in low-resource settings. One strategy for making training more scalable in low resource settings includes brief measures of treatment fidelity assessed from the perspective of the therapist or patient (Hogue et al., 2013). Therapist assessments of their own treatment fidelity can help them remain focused on the model's important therapeutic techniques and facilitate supervision.

Several attempts to rate treatment integrity from the therapist's perspective have been disappointing, often demonstrating poor correspondence between therapists' self-reports of fidelity and expert observer ratings, as well as overestimates of the use of prescribed techniques by therapists (Carroll et al., 1998; Hurlburt et al., 2010; Martino et al., 2009; Miller et al., 2004). However, other more recent investigations have had greater success in developing reliable and valid therapist self-reports of fidelity consistent with expert observers' evaluations (Caron & Dozier, 2022; Chapman et al., 2013; Hogue et al., 2014, 2015; McManus et al., 2012), suggesting that certain EBPs lend themselves to therapist or patient assessment of treatment fidelity more than other EBPs. In the case of BA, therapists and patients should be able to report accurately on the use of primary treatment techniques, such as activity scheduling, that rely on explicit discussions and use of forms to record such scheduling. There have been multiple previous studies to measure fidelity to BA from therapist self-reports (Kanter et al., 2015; Puspitasari et al., 2017). These studies rated global items (i.e., providing BA rational, reviewed homework, monitored activity) for presence versus absence. For these scales, the item pool was not developed using predictive validity methods and the item and rating scale language were not developed with input from community mental health stakeholders. No patient assessment of BA fidelity exists, though there have been attempts to have patients rate goal behaviors in the context of BA (Manos et al., 2010; Ryba et al., 2014).

Patient and therapist rated assessments of treatment fidelity could benefit from both the use of community stakeholders to develop items and a focus on predictive validity within the community setting. The inclusion of stakeholders to select and adapt items ensures that the final assessment includes items that are meaningful and relevant to therapists and patients. In developing brief assessments, it's also important to include items that represent the therapeutic techniques that lead to the treatment mechanisms ("targets") responsible for clinical benefit in the community settings. For example, when therapists focus on activity scheduling in BA sessions, does this lead to patients engaging in more of such activities? There is substantial evidence for the mechanism of action of BA in terms of the role of increases in activities and the reward value of activities driving changes in depression (Carvalho & Hopko, 2011; Christopher et al., 2009; Dimidjian et al., 2017; Gawrysiak et al., 2009), including a study conducted in a CMHC setting (Crits-Christoph et al., 2021). Accordingly, focusing a brief scale on the BA techniques driving change in the mechanism variables that are predictive of symptom improvement in the CMHC setting is the most effective way to create a theory-relevant fidelity measure that can help therapists focus on the techniques that lead to clinical benefit.

The goal of the current project was to conduct initial studies to develop parallel observer, therapist, and patient brief measures of BA fidelity, defining fidelity as the degree of utilization of a technique during a session. To develop these brief pragmatic measures to support the dissemination of BA in CMHC settings, our strategy was to select final items that occur with adequate frequency in actual community practice, are meaningful and clear to stakeholders, and are predictive of change in the important targets of BA driving symptom reduction in these settings. To ensure that our measures could also be useful for research purposes, we also selected items that covered the full range of facets of BA determined by experts, could be reliably measured by

independent raters, met standard psychometric reliability criteria, and provided the greatest amount of information.

General Methods

Overview

The research plan consisted of three initial steps: (1) a literature review followed by qualitative research methods to develop a comprehensive initial item pool relevant to community practice, (2) a tape-rating study of expert observer ratings of the BA fidelity item pool from a recent BA effectiveness study (Crits-Christoph et al., 2021) to reduce the item pool based on classical test theory supplemented by an item response theory analysis and an examination of the predictive validity of items, and (3) an initial prospective feasibility study to evaluate the acceptability and feasibility of rating BA fidelity from the perspective of patients and therapists.

Setting

The qualitative item review, effectiveness study used for observer ratings, and prospective feasibility study were conducted in partnership with Merakey, a large, private, nonprofit, CMHC that provides mental health and substance abuse services to primarily publicly-funded clients across multiple states. The specific CMHC where the studies were conducted employs approximately 80 mostly master's level psychotherapists and three to four psychiatrists, serving approximately 4,900 individuals per year (40% minorities). The most frequent diagnoses at the clinic include schizophrenia (39.8%) and MDD (33.9%).

Study 1: Development of the BA Fidelity Item Pool

Overview

To accomplish the generation of a comprehensive item pool and subsequent further review of items ("winnowing"), we followed steps for scale construction used in the Patient Reported Outcomes Measurement Information System (PROMIS) network funded by the NIH (Cella et al., 2007, 2010; DeWalt et al., 2007). These steps included the generation of a comprehensive initial item pool through a literature review as well as the generation of items by experts, binning and winnowing of items, and qualitative item review by stakeholders.

Method

Generation of Initial Item Pool

The investigators conducted a thorough review of the literature to ensure all existing instruments assessing fidelity to BA were found and evaluated. The literature beyond measures of adherence and competence alone was searched to include general quality measures utilized in diverse settings and fidelity measures that have successfully been rated as self-reports. Since existing measures of BA fidelity often included very general items not specifically focused on individual BA techniques, a thorough review of all materials detailing the specific techniques of BA was conducted, including published manuals and workbooks. The investigative team generated an initial set of BA fidelity domains and items to represent the facets of BA as included in the stand alone BA approach manualized by Martell et al., (2013). Item generation was also done by the BA trainer and supervisor of our prior BA effectiveness trial (Crits-Christoph et al., 2021) to ensure the item pool covered the important domains of BA as delivered in community settings. All items for the initial item pool were written from the perspective of the therapist completing the item. All relevant items, regardless of quality or redundancy, were retained at this step for further review.

Binning and Winnowing

Once a comprehensive list of BA fidelity items was generated, the items were given to members of a workgroup consisting of study investigators, an agency therapist/administrator, and the expert BA supervisor/trainer for binning and winnowing. The items were presented in preliminary bins (i.e., domains of BA techniques), but expert members of the workgroup were instructed to add or subtract bins as needed. Winnowing then occurred to eliminate any clearly inferior items. Members of the workgroup were asked to independently identify items that were inconsistent with each bin, redundant (i.e., they were asked to choose the best item for that bin), too narrow to be applicable to most sessions, too confusing, or not fully representative of BA as delivered in the community setting. In addition, members of the workgroup were asked to develop new bins if appropriate and to generate new items to cover the domain. The study principal investigator then reviewed the workgroup's ratings and created the revised item pool for qualitative item review, taking into consideration the suggestions of each workgroup member while also adapting each item to be accessible across literacy levels, unambiguous, of low cognitive difficulty, referencing the current session, and concise.

Qualitative Item Review

A qualitative item review (DeWalt et al., 2007) was conducted to ensure the set of items was comprehensive in measuring each domain, clear and understandable, and acceptable to respondents. The qualitative item review included both focus groups with therapists working in a CMHC setting and cognitive interviews with therapists and patients. A semi-structured interview for the focus groups was created by the principal investigator to explore therapists' opinions about the domains covered by the item pool and reactions to specific items. In addition, the focus groups were used to evaluate general opinions about the research partnership, familiarity with BA, usefulness of BA in community settings, usefulness of fidelity assessment, and methods for implementing the fidelity assessment to inform the development of a dissemination plan if the fidelity assessments were found to be feasible, acceptable, and effective at supporting training in BA.

We recruited participants for two focus groups with three therapists per group. The focus groups included three (50%) White therapists and three (50%) African American therapists. Of the six therapists, two (33%) identified as males and four (66%) as females. The first focus group included three therapists already trained in BA as part of our prior BA effectiveness trial (Crits-Christoph et al., 2021) while the second focus group included three TAU therapists with no formal experience with BA. We included both therapists already trained in BA as well as therapists not formally trained in BA to ensure that the language used for items, directions, and the rating scale was acceptable to the broad range of therapists working in the community mental health system, not just the therapists already trained by us who supported the BA training program. Our goal was to ensure that these fidelity assessments were acceptable from the start of training in community settings with stakeholders new to BA. The focus group audiotapes were transcribed and subsequently checked by a second individual. The principal investigator created the initial data dictionary for coding sessions using NVivo 11. The data dictionary creates a unique code, called a "node" in NVivo, for each question included in the semi-structured interview guide. Two independent judges used this code book to assign text from sessions into each node and to identify emerging themes using NVivo 11. Once complete, the judges met with the principal investigator to discuss differences and come to a consensus. Each judge then independently returned to the transcripts and reviewed/updated their coding. The final themes of the qualitative assessment were summarized by the principal investigator and used to develop a revised set of items in preparation for the psychometric study.

Finally, we recruited 11 patients who were receiving services for depression at the CMHC and 10 therapists delivering services at the CMHC for cognitive interviews (Cella et al., 2010) of the revised item pool. The therapists included seven (70%) White therapists and three (30%) African American therapists. Two (20%) identified as male and eight (80%) as female. Two of the 11 patients did not provide demographic data. The remaining nine patients included five (56%) White, three (33%) African American, and one (11%) Asian patient(s). Two (22%) identified as male and seven (78%) as female. We probed the thought processes engaged by persons responding to the measure to reveal sources of confusion or ambiguity in the instructions, item stems, response options, or time frame. Therapists provided feedback on the items from the focus groups that were written from the therapist's perspective and patients provided feedback on items adapted to be completed from the patient's perspective. For each item, the interviewer summarized any concerns discussed by patients and therapists. Stakeholders naïve to BA were included to ensure that item wording was acceptable to both community therapists starting BA training and community patients starting a new course of BA. The principal investigator used the results of the cognitive interviews to revise the pool of items for the therapist and patient assessments of BA fidelity.

Data Analysis

Interjudge agreement for coding of text from the focus groups into nodes and sub-nodes was assessed with Cohen's kappa (κ) computed using NVivo.

Results

Development of the Initial Item Pool

Based on the literature review by the investigative team and input from the BA expert, the initial BA fidelity item pool included 119 items (see Supplement 1). The following existing sources were used for item generation: (1) Quality of Behavioral Activation Scale (Q-BAS; Dimidjian et al., 2012 [unpublished]), (2) Collaborative Study Psychotherapy Rating Scale (Hill et al., 1992), (3) the BA workbook (Martell & Addis, 2004), and (4) the BA manual (Martell et al., 2013).

Reduction of Item Pool Based on Expert and Stakeholder Input

The binning and winnowing process reduced the initial list of 119 items to 47 items (See Supplement 2) across 6 rationally derived BA fidelity domains. These 47 items

were reviewed in the therapist focus groups. Six nodes were rated across the focus group transcripts with the fourth node, labelled "Review of BA Item Pool," broken down into eight sub-nodes (see Supplement 3 for structured interview guide). Interjudge agreement for selection of text was fair to good across nodes representing knowledge of research community partnership (κs across focus groups ranged from .53 to .62), familiarity with BA (κs = .56 – .70), usefulness of BA fidelity assessment (κs = .33 – .71), and methods for administration of BA fidelity assessment (κs = .78 – 1.00). The node representing usefulness of BA for CMHC settings had poor interjudge agreement (κs = .22 – .39) due to stakeholders having little to say in response to this question thus leading to limited data to code. The eight sub-nodes in the "Review of BA Item Pool" node represented specific questions regarding the BA fidelity item pool ranged from poor to good interjudge agreement (κs = .29 – .95). The sub nodes that represented questions probing for item clarity and wording to capture intensity had poor agreement across focus groups (κs = .08 – .57), again this is due to limited data rated in that node. Interjudge agreement was good for at least one focus group for sub nodes pertaining to questions about item pool facet coverage, relevance of items, length of items, item redundancy, changes to item wording, and identification of best items (κs = .29 – .95).

The themes extracted from the focus group ratings are detailed in Supplement 4. Stakeholders across focus groups felt that the item pool covered the relevant domains of BA, however, they did not find items probing commitment to treatment as relevant to community practice. Therapist stakeholders without prior experience in BA did not find items focused on introduction of BA model, homework, and nonspecific therapeutic techniques relevant to BA practiced in the community. Since these facets were acceptable to therapist stakeholders previously trained in BA, we retained items to address these facets. Therapists across both focus groups felt that the optimal length for a BA fidelity assessment completed after each session would be 10 to 12 items, taking no more than two to five minutes. In addition, therapists suggested individual items be brief, containing approximately 10 words. Overall, therapists agreed that items were clear. Therapist stakeholders were helpful in identifying redundant items as detailed in Supplement 1. Therapists also identified word phrases, such as "not in the mood," "fighting depression," "when you don't feel like it," and "problems," that they felt were invalidating to the patient's experience. Items were selected either avoiding these phrases or adapting the wording. We considered including phrases such as "throughout the sessions" or "thoroughly" within each item to better capture the intensity of each BA intervention, however, the therapists preferred item brevity with intensity encompassed in the rating scale anchors. Finally, therapists identified items that they felt best represented the BA techniques.

The focus group results were used by the investigative team to reduce the item pool to 32 items across six BA fidelity domains and to adapt item wording where needed (see Table 1). These 32 items were next evaluated in individual stakeholder cognitive interviews for clarity of item wording and content. Items where at least three participants in the cognitive interviews raised concerns regarding the content or wording are identified in Table 2. For the most part, few items were identified by multiple patients or multiple therapists as problematic regarding wording or content. These ratings were used in conjunction with psychometric analyses in Study 2 to select the final BA fidelity assessment items and adapt wording for final items if needed.

Table 2 provides interrater reliability for items, mean (*SD*), corrected item-total correlations (based on the total of 32 items), and partial correlations of the 32 items with improvement in the BADS Avoidance and BADS Activation subscales. In addition, items identified as problematic by patients or therapists in the cognitive interviews in terms of item content and wording are indicated. Table 2 also shows the slope parameters based on a unidimensional GRM IRT model fitted to the 32 items. Also shown in Table 2 is the final decision about each item, giving priority to the predictive validity of each item and coverage of the facets of BA but also taking into account any concerns regarding variability and feedback from patients and therapists about the content and wording of items

General Themes from Focus Groups

For nodes targeting stakeholders' perspectives of the research community partnership, familiarity with BA, usefulness of BA in the community, usefulness of BA fidelity assessment, and methods of administration, major themes that were evident from this qualitative analysis included: (1) overall, therapists were enthusiastic about research conducted in community settings that had the potential to include the perspectives of minority stakeholders; (2) therapists previously involved in our research program had a generally positive attitude toward the academic community partnership; (3) despite a long-standing partnership, therapists in this large outpatient community setting who were not directly involved in one of our studies were unaware of the partnership and expressed ambivalence about using a treatment manual and using a treatment with time limits; (4) therapists who were not trained in BA through our research program felt that they were already using aspects of BA in the clinic, but they reported little formal training in BA as a stand-alone treatment and had limited knowledge of the breadth of techniques within BA; (5) although agreement

Table 1 Reduced BA Fidelity Item Pool with Domains

STARTING BA/TEACHING THE GENERAL BA MODEL
1. The therapist reviewed the goal of BA: to break the cycle of depression by trying new activities.
2. The therapist described how his/her job was to guide the client with strategies to become more active.
3. The therapist talked about how changing what one does can change how they feel.
4. The therapist asked the client for feedback after describing the BA model.
GENERAL STRUCTURE/HOMEWORK
5. The therapist followed an agenda in the session.
6. The therapist reviewed the activity and mood monitoring chart.
7. The therapist and client decided on specific homework for the next session.
8. The therapist reviewed the client's homework in the session.
9. The therapist reviewed the activity schedule to see what made the past week better and what made it worse.
NONSPECIFIC THERAPEUTIC TECHNIQUES
10. The therapist was warm throughout the session with the client.
11. The therapist was nonjudgmental throughout the session with the client.
12. The therapist was encouraging with the client throughout the session.
TECHNIQUES EXPLORING AVOIDANCE
13. The therapist reviewed how avoidance is an understandable response to sadness, but can make depression worse.
14. The therapist and client talked about the many forms of avoidance (i.e., isolating oneself, ruminating, overeating, drug use, and/or excessive screen time).
15. The therapist and client explored alternatives to rumination (i.e., distraction and mindfulness).
16. The therapist reviewed how it is important to act even when not feeling motivated.
17. The therapist and client explored how rumination leads to lower mood.
18. The therapist encouraged the client to shift to problem-solving rather than relying on avoidance.
ACTIVATION TECHNIQUES
19. The therapist encouraged the client to monitor their activities and mood next week.
20. The therapist and client scheduled specific activities to try this week.
21. The therapist and client planned activities that would increase pleasure.
22. The therapist and client talked about establishing routines.
23. The therapist focused on planning activities that could lead to a sense of accomplishment.
24. The therapist and client discussed pros and cons of specific actions.
25. The therapist and client identified barriers to being active.
26. The therapist focused on identifying long-term goals and aspirations.
27. The therapist and client broke down difficult assignments into smaller tasks.
28. The therapist helped the client schedule one or more activities in detail, including what, where, when, and with whom.
RELAPSE PREVENTION
29. The therapist reviewed how the client has improved avoidance, rumination, isolation, etc.
30. The therapist and client discussed possible future problems and how to tackle them.
31. The therapist reviewed the strategies that have been helpful for the client.
32. The therapist and client planned what the client could do to prevent a relapse.

among coders was not adequate for the probe into usefulness of BA in community settings, individual statements identified by at least one coder representing the thoughts of at least one stakeholder who was not previously trained by us in BA indicated questions regarding the feasibility of BA homework in the community setting, doubts about the psychoeducation that is included in BA, and thoughts that BA lectured too much to the patient; (6) in contrast, therapists previously trained by us in BA were more positive about the treatment and said they used it with other non-study clients; (7) therapists felt that a BA fidelity self-report would help them track what they did from session to session and prepare for future sessions; and (8) therapist stakeholders preferred tablet administration to paper administration but had concerns about technology support in community settings and time burden.

Discussion

We developed a large comprehensive item pool to ensure a thorough evaluation of the most useful items for assessing BA fidelity from the therapists' and patients' perspective to support the dissemination of BA for the treatment of MDD in community mental health settings. We then reduced the item pool by first using a team of experts for binning and winnowing and then including community clinicians for fine-tuning the items. This process demonstrates

Table 2 *Psychometrics, Descriptive Statistics, Predictive Validity, and Stakeholder Feedback from the Cognitive Interviews*

Item	ICC (3,3)	*M*	*SD*	Item-Total r	IRT Slope	Cognitive Interview Concerns	BADS Avoidance		BADS Activation		Decision
							M 1 r_p	M 2/3 r_p	M 1 r_p	M 2/3 r_p	
STARTING BA/TEACHING THE GENERAL BA MODEL											
1	0.91	1.84	0.97	0.50	1.63	none	−.35*	−.05	.14	−.18	retain
2	0.82	1.39	0.64	0.30	1.17	therapist	−.29*	−.21	.05	−.23	eliminate low item-total r
3	0.74	1.83	0.73	0.59	2.04	none	−.36*	−.08	−.00	−.19	eliminate focus groups redundant with 1
4	0.63	1.29	0.41	0.45	1.71	patient	−.31*	−.14	−.14	−.22	eliminate low item-total r, patient concerns wording
GENERAL STRUCTURE/HOMEWORK											
5	0.86	2.59	1.17	0.61	1.90	patient	.02	.07	.32*	.16	eliminate patient concerns wording
6	0.81	1.93	0.88	0.72	3.42	none	.14	.38*	.40*	.18	retain
7	0.84	2.71	0.85	0.76	2.67	none	.00	.17	.34*	.00	eliminate predictive validity not as strong as item 6
8	0.89	1.88	0.93	0.65	2.14	none	.03	.25	.29*	.23	eliminate low predictive validity
9	0.79	1.69	0.81	0.67	2.60	none	−.02	.30	.26	.21	eliminate predictive validity not as strong as item 6
NONSPECIFIC THERAPEUTIC TECHNIQUES											
10	0.71	3.65	0.50	0.26	0.53	therapist	−.11	−.22	.11	−.12	eliminate low item-total r, IRT slope, therapist concerns wording
11	0.63	3.63	0.51	0.44	0.98	therapist	−.26	−.21	.17	−.09	eliminate low item-total r, therapist concerns wording
12	0.62	3.59	0.57	0.55	1.16	none	−.33*	−.07	.08	.05	retain
TECHNIQUES EXPLORING AVOIDANCE											
13	0.66	1.53	0.69	0.54	1.45	none	−.06	.09	.28*	.01	retain
14	0.72	1.66	0.64	0.58	1.62	patient	−.01	.00	.12	−.01	retain adapt wording
15	0.59	1.32	0.51	0.37	0.49	patient	−.39*	.02	−.05	.12	eliminate low ICC, item-total r, IRT slope
16	0.73	1.42	0.61	0.48	1.80	patient	−.21	−.02	−.04	−.13	retain adapt wording
17	0.61	1.33	0.46	0.20	0.27	patient	−.32*	−.03	−.03	.12	eliminate low item-total r, IRT slope
18	0.59	1.50	0.57	0.23	0.53	patient	−.07	−.13	−.10	−.07	eliminate low item-total r, IRT slope
ACTIVATION TECHNIQUES											
19	0.89	2.38	1.04	0.72	2.72	none	−.18	.03	.31*	−.04	eliminate focus groups redundant with 20
20	0.86	1.96	0.88	0.62	2.36	none	−.08	.07	.39*	.15	retain
21	0.83	1.74	0.77	0.58	1.96	therapist	−.05	.06	.27	.10	retain modify wording
22	0.63	1.17	0.35	0.30	0.91	none	−.07	−.09	.32*	−.07	eliminate low item-total r
23	0.75	1.55	0.59	0.58	2.28	therapist	−.06	.05	.44*	.22	retain combined with item 21 based on qualitative results
24	-	1.35	0.32	0.03	-0.03	none	.05	−.08	−.15	−.01	eliminate low ICC, item-total r, IRT slope
25	0.71	1.55	0.57	0.45	1.38	patient	−.31*	−.05	−.11	−.15	eliminate low item-total r, patient concerns wording
26	0.58	1.72	0.61	0.17	0.39	none	.06	.04	.15	.04	eliminate low ICC, item-total r, IRT slope
27	0.53	1.33	0.46	0.30	0.70	none	−.13	−.11	.13	.00	eliminate low ICC, item-total r, IRT slope
28	0.85	1.57	0.72	0.56	2.51	none	−.06	.04	.38*	.02	retain
RELAPSE PREVENTION											
29	0.51	1.17	0.33	0.31	0.98	none	−.14	−.01	.17	.22	eliminate low predictive validity
30	-	1.31	0.29	0.17	0.29	none	−.12	−.05	−.08	−.04	eliminate low ICC, IRT slope, predictive validity
31	0.35	1.41	0.42	0.47	0.94	none	−.23	.11	.42*	.20	retain, best predictive validity in facet
32	0.10	1.06	0.16	0.18	0.53	none	−.22	.11	.17	.42*	eliminate low ICC, item-total r, IRT slope

Note: ICC = Interclass Correlation Coefficient; IRT = Item Response Theory. For Cognitive Interview Concerns, "none" represents only minor concerns from few participants while "therapist" or "patient" represents concerns raised by at least 3 participants of each stakeholder group. * $p \leq .05$; M1 represents partial correlation between item and month 1 BADS score controlling for baseline; M 2/3 represents partial correlation between item and average of months 2 and 3 BADS scores controlling for baseline

the richness that results from the inclusion of a broad range of stakeholders in the selection and adaptation of items for

use in community settings. While experts were extremely helpful in fleshing out a comprehensive set of items covering the important facets of BA as defined by Martell and colleagues (2013), community clinicians were particularly important for identifying items that best resonated with the use of BA in community settings and for adapting wording to best meet the needs of community patients.

Not only were the focus groups particularly informative regarding the selection of items for our BA fidelity assessments, they were also efficiently used to gather important information to support planning of future dissemination of the BA fidelity assessments. The emerging differences in perspectives between clinicians who were versus who were not previously trained by us in BA supported that future implementation of both BA and the BA fidelity assessment would need to continue to include strong initial socialization to the interventions to engage community clinicians.

Study 2: Tape Rating Study of BA Fidelity Items

Overview

The reduced pool of 32 items developed in Study 1 via expert and stakeholder input was evaluated for both psychometrics and predictive validity. Our goal was to select a final limited number of items to create a scale that was reliable, valid, and predictive of the change in BA targets that drive symptom improvement. For this study, each of the 32 items was translated into observer rated versions and was rated by trained independent observers on BA sessions collected as part of a BA pilot effectiveness trial in a community mental health setting (details presented in Crits-Christoph et al., 2021). A brief description of the patients, treatments, and raters is provided below.

Methods

Patients

We recruited patients (aged 18 to 70) from those seeking services at the CMHC. All adult patients at the clinic were administered the Quick Inventory for Depressive Symptomatology (QIDS; Rush et al., 2003) at intake. If a patient scored at least 11 on the QIDS, was able to read English at the fourth grade level, and was interested in participating in the study, they were scheduled for a study baseline assessment after providing informed consent. Patients were included in the study if they were diagnosed with MDD by trained diagnosticians at their baseline assessment. Exclusion criteria were: (1) a diagnosis of bipolar disorder; (2) current or past diagnosis of schizophrenia, psychosis, MDD with psychotic features, or seizure disorder; (3) depression due to organic pathology; (4) symptoms of substance/alcohol abuse requiring immediate referral to substance abuse treatment; (5) problems requiring immediate referral to a partial hospitalization program; or (6) suicidal ideation judged by the clinic to require services more intensive than once per week psychotherapy.

The patient sample included in this study consisted of 64 patients who participated in the training phase or randomized phase of the Crits-Christoph et al. (2021) BA pilot effectiveness trial, received at least one session of BA, and had at least one post-baseline assessment using the Behavioral Activation for Depression Scale (BADS; Kanter et al., 2007). This sample was aged 18 to 70 ($M = 40.29$, $SD = 12.63$) and consisted of 50 (78%) females. The racial/ethnic breakdown for the 56 patients who self-reported their racial group was 22 (39%) white, 27 (48%) black, and 7 (13%) other or mixed race, with 6 (9%) identifying as Hispanic.

Raters and Sessions

We recruited and trained four advanced graduate student judges with prior clinical training in BA. An early session from each patient in the BA condition of our pilot effectiveness trial (Crits-Christoph et al., 2021) was rated independently by three judges using a balanced incomplete block design. We selected an early session of BA in order to conduct ratings on all patients included in the sample. In this community administration of BA, therapists were trained to incorporate techniques to introduce the treatment, assign homework, and begin planning activities from the start in session 1. Session 3 was used when available to sample sessions that included the full range of BA techniques. We substituted session 2 or 1 if session 3 was unavailable. Session 2 was sampled for 11 cases (17%) and Session 1 was sampled for 11 cases (17%). Judges rated each BA fidelity item on a 1 (*not at all*) to 5 (*thoroughly*) scale. Judges initially participated in a series of five training sessions to review the BA fidelity items and rate practice sessions. For each practice session, judges first rated items independently, discussed disagreements, and finally reached a consensus judgement. Monthly recalibration sessions with the judges were conducted during the time that final ratings were made.

Treatment

Therapists delivered three months of weekly sessions of BA following the BA manual by Martell et al., (2013) modified by Christopher Martell to fit a 9-session format. Behavioral activation focuses primarily on identifying activities and

contexts that are reinforcing and align with one's long-term goals. Specific strategies that BA uses are: (a) self-monitoring, (b), planning daily activities, (c) evaluating the degree of pleasure and accomplishment experienced during engagement in specific activities, (d) exploring different ways to achieve individuals' goals, (d) targeting certain behavioral deficits through role playing, (e) assessing and treating avoidance behaviors, (f) creating and adhering to routines, and (g) targeting rumination with different behavioral techniques, such as redirecting attention away from ruminative thoughts towards direct and immediate experiences. Participants' weekly homework included daily monitoring of mood and activity. Some patients could also be asked to complete a worksheet to identify triggers, responses, avoidance patterns, and consequences of these experiences.

Target Assessment

The BADS (Kanter et al., 2007), a self-report scale of 25 questions, was rated on a 7-point scale ranging from 0 (*not at all*) to 6 (*completely*) to assess BA targets. The scale has demonstrated adequate internal consistency, construct validity, and predictive validity (Kanter et al., 2007, 2009). The current study focused on two subscales of the BADS that measured increased activity and decreased avoidance, respectively.

Data Analysis

Selection of the final item pool was guided by a comprehensive evaluation of psychometrics and stakeholder input on each item. In order to develop a final item pool that was reliable and valid for use in observer-, therapist-, and patient-rated formats, interrater reliability was calculated for each of the 32 items using Shrout and Fleiss's (1979) intraclass correlation coefficient ICC(3,3) with a standard of 0.60 used to identify items that could be reliably rated by independent judges. The ICC(3,3) was also calculated for the mean across the final items selected to inform the reliability of the final BA fidelity score. We calculated descriptive statistics and corrected item-total correlations for each item with the goal of selecting items that had corrected item-total correlations of 0.50 or greater.

We also examined the items using a unidimensional Item Response Theory (IRT) method to further evaluate items using the SAS (version 9.4) IRT procedure. We implemented the Graded Response Model (GRM), which is appropriate for analyzing the polytomous Likert-style item responses used in the BA fidelity assessment. IRT assumes the amount of information that each item provides is not evenly distributed across the entire continuum of the latent construct. The slope parameter was used to quantify the amount of information provided by each item, with a slope value below 1.0 used as a threshold for detecting less discriminating items (Embretson & Reise, 2000).

Predictive validity analyses were further examined to select items that were predictive of change in BA targets. Session 3 observer ratings of BA fidelity were used as predictors of change in the avoidance and activity subscales of the BADS. The BADS outcomes are examined in terms of short-term change (baseline to month 1) and longer-term change (baseline to the average of months 2 and 3). The average of the months 2 and 3 assessments was used as an index of longer-term change due to attrition. Partial correlations of the BA fidelity items with BADS scores at month 1 (and separately with the average of months 2 and 3) were computed, controlling for the respective BADS baseline score. Items with a medium effect (r of 0.30 or greater) with one or both of the two target measures were a priority for retention in the final item pool.

We used an exploratory factor analysis with varimax rotation to evaluate whether the final BA fidelity items resulted in a factor structure consistent with the original rationally derived domains. Note that our goal was to develop a very brief assessment that could be efficiently used as an intervention to support training rather than to have a comprehensive assessment with subscales that fully covered our original rationally derived BA fidelity domains. Our priority was to select items that were predictive of target change so that the assessment could focus clinicians specifically on those techniques that could lead to clinical benefit. For this reason, we conducted an exploratory factor analyses to evaluate whether our final brief assessment broadly covered the domains. We also computed internal consistency for the items included in the final BA fidelity assessment using Cronbach's α.

Results

Overview

The final item pool was selected to include items that were reliable, occur with adequate frequency and variability across sessions, were not redundant with other items, provide a meaningful amount of information, and were predictive of target change. Our goal was also to incorporate stakeholder opinions, gathered through focus groups and cognitive interviews, into the selection of items and the size of the final item pool. Stakeholders in our qualitative evaluations agreed that a brief measure would be most useful and uniformly agreed on the facets covered by our item pool. We therefore selected 10 items that best met our selection criteria, including one item to represent the "starting BA"

facet, one item representing the "homework" facet, one item representing the "nonspecific therapeutic techniques" facet, three items representing the "avoidance techniques" facet, three items representing the "activation techniques" facet, and one item representing the "relapse prevention" facet, in order to have a brief but comprehensive assessment that also emphasized the primary focus of BA on techniques to address activation and avoidance.

Item Selection

Items 1–4 were evaluated for the "starting BA" facet. Items 1 and 3 met selection criteria for interrater agreement, item-total correlation, IRT slope, patient and therapist feedback on wording and content, and predictive validity. Item 1 was selected to represent this facet based on its excellent interrater reliability. Items 5–9 were evaluated as part of the "homework" facet. Items, 6, 7, and 9 all met selection criteria for interrater reliability, item-total correlations, IRT slope, and patient and therapist feedback. Item 6 was selected because of strong predictive validity. Items 10–12 were evaluated as part of the "nonspecific therapeutic techniques" facet. Only item 12 met all selection criteria, so it was selected for the final item pool.

Items 13–18 were evaluated for the "avoidance techniques" facet. Items 15, 17, and 18 were eliminated due to low item-total correlations and IRT slope coefficients less than 1, indicating that the items did not adequately discriminate among cases. The three remaining items were selected based on adequate interrater reliability, item-total correlations, and IRT slopes, although none demonstrated moderate effects for predictive validity. Cognitive interviews indicated patient concerns with the wording of items 14 and 16, resulting in modifications to item wording in the final item pool. For item 14, the phrase "isolating" was changed to "staying away from people" and the phrase "excessive screen time" was changed to "too much time with electronics." For item 16, the phrase "not feeling motivated" was changed to "feeling tired, stressed, or depressed."

Items 19–28 were evaluated for the "activation techniques" facet. Items 19, 20, 21, 23, and 28 met initial selection criteria. Item 19 was eliminated because therapists in the focus groups originally saw items 19 and 20 as redundant. Furthermore, items 21 and 23 were seen as highly redundant by therapists since both probed for planning of rewarding activities. Therapists in the qualitative evaluation saw the items as redundant, with activities that would "increase pleasure (item 21)" and "lead to a sense of accomplishment (item 23)" both representing types of rewarding experiences targeted by BA activation techniques. Therapists suggested combining these two items, so item 21 was selected for the final item pool but was modified to probe for planning activities that increased pleasure or led to a sense of accomplishment.

Finally, items 29–32 were evaluated for the "relapse prevention" facet. None of the four items fully met our selection criteria. However, item 31 was retained because it had an item-total correlation and IRT slope coefficient only slightly below the selection criteria and strong predictive validity. In addition, it was determined that the interrater reliability was lower than the selection criteria because it represented relapse prevention techniques that were not common in the sessions rated early in treatment. Since this facet of BA was judged important for tracking the fidelity of BA across treatment, recognizing that it would not occur early in treatment, the item was retained to ensure a comprehensive set of final BA fidelity items.

Final BA Fidelity Assessment

The 10 items selected for the final BA fidelity assessment are presented in Supplement 5, including the minor modifications to wording described above for original items 14, 16, and 21. Based on an analysis of these 10 items (prior to final wording modifications) from the independent observer ratings, all items demonstrated good corrected item-total correlations (ranging from 0.40 to 0.66 based on the 10 item total).

All 10 items had adequate IRT slope coefficients indicating that the items provided adequate information to differentiate respondents. We inspected the item information curves from the IRT analyses for each of the 10 items and found that curves for eight items were not flat at any region, suggesting that the items were reliable across the range of the latent variable. Original items 12 and 31 did reveal some flattening of the curve and both had slope coefficients close to 1. These items were ultimately selected because they demonstrated good corrected item-total correlations and predictive validity.

The exploratory factor analysis resulted in three factors with eigenvalues greater than 1.0 that together accounted for 72% of the variance. The three items selected to represent specific activation techniques (original items 20, 21, 28) loaded highest on the first factor. The three items selected to represent techniques focused on avoidance (original items 13, 14, 16) as well as the single item selected to represent teaching the BA model (original item 1) loaded highest on the send factor. The items representing techniques focused on homework (original item 6), nonspecific therapeutic techniques (original item 12), and relapse prevention (original item 31) loaded highest on factor three.

The total score of the 10 items, representing overall fidelity to BA techniques, demonstrated high internal consistency ($\alpha = .83$) and interjudge reliability ($\kappa = .88$). The

overall mean (*SD*) for the average of the 10 items was 1.87 (*.46*). As an initial evaluation of the predictive validity of the 10-item total score, we computed partial correlations between the total score rated on session 3 by observers and the patients' ratings of activation and avoidance from the BADS. The total score significantly predicted short-term change on the BADS activation scale, $r_p = .40$, $p = .004$, but not longer term change, $r_p = .053$, $p = .705$. The total score, rated by observers at session 3, did not significantly predict the short-term, $r_p = -.19$, $p = .187$, or longer-term, $r_p = .10$, $p = .455$, patient ratings of activation. The activation subscale, composed of the 3 activation items, and the avoidance subscale, composed of the 4 items that loaded on the second factor, demonstrated high internal consistency ($\alpha = .93$ and $\alpha = .79$, respectively) indicating that these subscales may be used to assess the specific activation and avoidance techniques included in the BA model. Since factor 3 represented a diverse group of single item facets we recommend a focus on the total score or activation and avoidance subscales to meet the unique clinical and research needs.

Discussion

Building on our comprehensive qualitative approach for developing the initial item pool described in Study 1, Study 2 implemented a comprehensive quantitative approach for the final selection of BA fidelity items based on expert observer ratings. Responding to the perspective of stakeholders involved in our qualitative analyses, we selected a brief set of only 10 items that best represented the range of BA techniques relevant to community practice. We selected items that reliably represented the important domains of BA fidelity as delivered in the CMHC setting but also focused on items that could be reliably assessed by independent observers so that we could have parallel measures from the perspective of therapists, patients, and independent observers to support flexible administration of the BA fidelity assessment across both clinical and research settings. IRT analyses informed the selection of items that provided a meaningful amount of information that is particularly important for such a brief instrument. Finally, our focus on the predictive validity of items suggests that our final assessments will have utility in assessing fidelity to specific techniques that drive change in BA and lead to clinical benefit in the CMHC setting.

The exploratory factor analysis provided preliminary evidence that our final 10-item assessment broadly represented the original rationally derived domains of this stand-alone model of BA described by Martell et al., (2013). Like most comprehensive fidelity scales, the total score of the 10-item assessment represents the full range of BA techniques and can be rated high based on different techniques consistent with the phase of treatment. Initial sessions of high fidelity BA may receive relatively higher ratings of fidelity based on the use of techniques to introduce the model, build a working relationship, and educate the patient. Although the bulk of sessions should include a specific focus on activation techniques, later sessions where activation schedules have already been mastered may focus more heavily on relapse prevention. For this reason, the total score should be interpreted as representing general fidelity to the overall BA model. Our initial results demonstrated predictive validity for the 10-item scale in relation to early change in activation as rated by the patient. Future research will need to examine the full predictive validity of the tool as rated by therapists, patients, and observers.

The factor analysis results included an activation factor and avoidance factor representing the primary techniques within this BA model that may have utility for assessing these specific facets of BA. While the total 10-item score may be useful as an assessment of overall fidelity to the broad model of BA that includes a variety of specific and nonspecific facets, the activation and avoidance subscale scores can be used as reliable assessments of these specific facets of BA. For training purposes specifically, it will be important for the therapist and supervisor to monitor both overall fidelity to the BA model that includes nonspecific therapeutic techniques, as well as fidelity to the specific activation and avoidance facets that are central to BA. Our results also included a single factor representing other BA facets covered by our assessment. Given our a priori choice to include only single items from each of these 4 facets, we did not expect individual factors representing each of these domains. Future research may evaluate the validity of the specific activation and avoidance factors for assessing specific techniques included in other models of BA or in implementations of activation techniques as a single component of a broader treatment approach.

Our goal was to develop a brief single item pool that could serve as the basis of pragmatic therapist and patient rated measures to support training in community settings as well as an independent observer version that could be used to quantify fidelity to BA for both clinical and research purposes. Our approach was to strike a balance between reliability, content validity, and predictive validity to optimize the validity of the item pool for both training and research. Standard psychometric and IRT analyses were used to select reliable items to meet both our clinical training and research goals. To ensure that we selected only items that could be meaningfully assessed from the therapist's perspective to support clinical training, we used predictive analyses based on early session recordings to select items that were predictive of changes in the important mechanisms of BA. It is

possible that items representing other techniques that usually occur with greater frequency later in treatment would have greater predictive validity if rated beyond session 3. For this reason, we included predictive validity as only one selection criteria. Although the items selected for our "starting BA", "homework", nonspecific techniques", and "activation" facets demonstrated predictive validity, none of the items representing the "avoidance techniques" facet met our criteria for predictive validity. To maintain a valid item pool that could be used to evaluate the full range of BA techniques that could occur across treatment we selected avoidance items that met our other selection criteria.

Study 3: Prospective Study of Feasibility and Acceptability of Therapist and Patient Assessments of BA Fidelity

Overview

We conducted a preliminary assessment of the feasibility and acceptability rating the final BA fidelity items from the perspective of both therapists and patients (see Supplement 6). The final pool of 10 items selected and modified from Study 2 were translated to both patient and therapist perspectives and administered following sessions of BA delivered in a CMHC setting.

Methods

Patients

Thirty patients seeking services for depression at the same CMHC where the pilot effectiveness study (Crits-Christoph et al., 2021) was conducted were recruited to participate in the feasibility sample. All patients were screened using the QIDS (Rush et al., 2003). Patients were eligible if they scored 11 or above on the QIDS and were 18 and older. The exclusion criteria were: (1) current substance abuse or dependence requiring primary referral to substance abuse program, (2) any current or past psychotic disorder, and (3) significant suicidal risk/ideation. These exclusions were specified due to safety and to not distort normal functioning of the clinic. Patients were paid $25 for the baseline assessment and $25 for the acceptability assessment.

Of the 30 recruited patients, a total of 22 patients consented to participate in the study and had at least one session of BA. Demographic data were only available for 21 patients. This sample included six (29%) white, 11 (52%) African American, and four (19%) patients who were other or mixed race. Of these patients, four identified as (19%) male and 17 (81%) as female. Ages ranged from 18 to 59 with an average age of 35.75 ($SD = 13.31$).

Therapists

Five therapists were recruited from those providing services at the CMHC and none had previously participated in our study of the feasibility of BA. All therapists were provided with the manual for BA and participated in a 1-day workshop on BA delivered by the same BA expert that provided training as part of our BA feasibility trial. Therapists received an honorarium of $100 for every three patients treated in the feasibility study as well as $25 for completion of the acceptability measure. Therapists participating in the feasibility study of the BA fidelity assessments included three (60%) White and two (40%) African American individuals – of these five therapists, two (40%) identified as male and 3 (60%) as female.

Treatment

Therapists delivered three months of weekly sessions of BA following the same general BA manual by Martell et al., (2013) modified by Christopher Martell for the prior pilot effectiveness study (Crits-Christoph et al., 2021) to fit a 9-session format.

BA Fidelity Assessment

Therapists and patients completed their respective BA fidelity assessments on handheld tablet computers directly following each session. For both patient and therapist versions, each item was rated on a 1 (*not at all*) to 5 (*thoroughly*) scale.

Ratings of Acceptability

Patients rated the acceptability of the BA fidelity assessment, including ratings of overall satisfaction, time burden, length, and comprehension of items. All patients in the feasibility sample completed the acceptability measure following session 3. Therapists rated the acceptability of the BA fidelity measure in terms of overall satisfaction, length, understandability of items, relevance to practice of BA in the CMHC setting, usefulness following BA training, helpfulness for monitoring BA, usefulness after each BA session, likelihood of recommending to other therapists interested in using BA, and comprehensiveness of item coverage. Therapists completed acceptability ratings after treating three cases in the study.

Feasibility Assessment

The feasibility of using the assessments was measured as the percentage of sessions that the fidelity assessment was completed by therapists and by patients.

Data Analysis

Using the patient and therapist versions of the scale collected as part of the feasibility study, we calculated preliminary internal consistency estimates. We also calculated descriptive statistics for acceptability ratings across therapists and patients to evaluate whether stakeholders had a generally positive attitude towards the utilization of our assessments of BA fidelity. Finally, descriptive statistics of the feasibility criteria were calculated.

Results

The 10-item therapist and patient versions of the BA fidelity assessment were evaluated in a feasibility study conducted in the CMHC setting. Twenty patients completed the acceptability assessment following session 3. One patient indicated that they did not wish to continue in therapy or complete any further study assessments and was not contacted for an acceptability assessment. A second patient could not be reached to complete the acceptability assessment. All therapists completed the acceptability assessment.

Data collection was interrupted by the COVID-19 pandemic. At the time the study was terminated because of COVID-19, 22 (100%) patients had completed at least one BA fidelity assessment on a tablet computer following a BA session. In addition, therapists had completed the BA fidelity assessment at least once for each patient. Thirteen (59%) patient participants were still engaged in treatment as part of this study when the study was terminated. These 13 patients had attended between 1 and 7 sessions of BA.

Patient ratings of acceptability were uniformly high across the 20 patient participants who completed the acceptability assessment, with a mean rating of overall satisfaction of 4.8 ($SD = 0.41$) on the scale from 1 (*not at all satisfied*) to 5 (*extremely satisfied*). Mean ratings for the other satisfaction items were also high: length ($M = 4.3$; $SD = 1.12$), time burden ($M = 4.7$; $SD = 0.73$), comprehension of items ($M = 4.8$; $SD = 0.41$). Though based on a small sample (n = 5), therapist mean ratings of acceptability were also high: overall satisfaction ($M = 4.6$; $SD = 0.89$), length ($M = 4.8$; $SD = 0.45$), understandable ($M = 4.8$; $SD = 0.45$), relevant to practice of BA in setting ($M = 4.8$; $SD = 0.45$), useful following BA training ($M = 4.8$; $SD = 0.45$), helpful to monitor BA ($M = 4.8$; $SD = 0.45$), useful after each BA session ($M = 4.4$; $SD = 0.89$), recommend to other therapists ($M = 4.6$; $SD = 0.55$), and comprehensive in covering range of BA techniques ($M = 4.6$; $SD = 0.89$).

Feasibility of completing the BA fidelity assessment was good in the CMHC setting. Only sessions attended by participants prior to study termination due to COVID-19 were included. Across the 22 patient participants, BA fidelity assessments were completed following 90% of attended sessions by patients and 93% of attended sessions by therapists. Selecting the first assessment completed for each patient, the BA fidelity assessment as measured from the patient's perspective had good internal consistency ($\alpha = .93$) with an average score of 4.43 ($SD = 0.75$) on a scale ranging from 1 to 5. Selecting only the first assessment for each patient and adjusting ratings on each item by the therapist mean to account for patients nested within therapists, the BA fidelity assessment as measured from the therapist's perspective had good internal consistency ($\alpha = .94$) with an average score of 2.72 ($SD = 0.98$) on the scale from 1 to 5.

Summary and Concluding Discussion

We implemented qualitative and psychometric analyses, including predictive validity, to derive an initial item pool, refine it, and reduce it to a manageable set of 10 BA fidelity items. By starting with qualitative approaches to derive and adapt the item pool, we developed BA fidelity assessments that can be useful and meaningful for stakeholders delivering BA in CMHC settings. A comprehensive quantitative approach to the final selection of items was used to develop an efficient measure that is reliable, meaningful, and useful for monitoring the implementation of BA for both clinical and research purposes. The observer-rated version of the 10-item scale was found to have good interjudge and internal consistency reliabilities, and both the therapist and patient versions of the final BA fidelity assessment demonstrated good internal consistency and reliability. IRT analyses indicated that all 10 items provided adequate to good amounts of information.

This is the first psychotherapy fidelity scale we are aware of that has used predictive validity of target (mechanism) variables as an aid in constructing the scale. This method ties the scale to the theory-relevant patient variables that are hypothesized to change when the therapist implements relevant techniques. The scale avoids inclusion of items that, while thought to be important by treatment manual developers, have little direct impact on the processes the therapy is designed to change. It is possible that techniques excluded from the scale due to failed predictive validity of short-term targets may be important to long-term change. However, in the case of BA delivered in the CMHC setting, an important goal of treatment is to provide relatively fast activation and

the techniques that further such early improvements should be prioritized. Focusing on rapid improvement is especially relevant in the CMHC environment where the median number of sessions attended is 5 (Connolly Gibbons et al., 2011).

The current study only provides very preliminary information on the patient and therapist versions of the BA fidelity assessment. Both patients and therapists completed the BA fidelity assessments using tablet computers following weekly therapy sessions. Ratings of acceptability were high for both patients and therapists who used the respective brief BA fidelity assessments in a community mental health setting. Importantly, it was feasible for patients and therapists to complete the scale following the majority of sessions. Both patients and therapists found the items understandable and relevant, reflective of our inclusion of community stakeholders in the development and adaptation of final items.

The BA fidelity assessments can be completed by clinicians, patients, and expert observers to meet the needs of unique clinical and research settings. Clinicians can both efficiently monitor their own use of BA techniques following training and review their ratings with their supervisors to support training. Clinicians and their supervisors may also find it useful to monitor patient self-reports of BA fidelity to have a full understanding of the perspective of the patient. This may highlight for clinicians that techniques are salient for the patient in the session and may indicate techniques that might need to be further emphasized with the patient. Finally, the observer rated measure can be used clinically to evaluate the effectiveness of training efforts and determine how best to use training resources to support BA implementation and for research purposes to evaluate fidelity to BA.

Our initial feasibility trial indicates that the BA fidelity assessments rated from the perspective of the therapists and patients in a CMHC setting are feasible and acceptable. Future research is needed to broadly evaluate the feasibility and acceptability of these assessments in community practice and to evaluate the accuracy of assessments rated from the perspective of therapists and patients. Further, future investigations should focus on evaluating the effectiveness of these therapist and patient ratings of BA fidelity at improving therapist fidelity to BA techniques following community workshops. Our preliminary results suggest that clinicians may find it useful to use these brief assessments to review sessions and support treatment planning for future sessions. Especially in community settings that lack the full resources to implement costly training and evaluation programs for new evidence-based treatments, it will be important to evaluate whether scaled back training programs that include some expert-led training activities supported by these efficient assessments can achieve adequate fidelity in community settings.

Limitations of the current program of research include a focus on stakeholders in CMHC settings that may not generalize to other clinical and research applications of BA. Although we started with a broad item pool, it is possible that alternative items not included might better capture the BA techniques relevant to community implementation. In addition, we selected items that demonstrated good predictive validity when rated by observers. Future investigations will need to evaluate the predictive validity of the items when rated by the therapist and patient. Our feasibility study indicated that patients rated fidelity of sessions higher on the 5-point scale ($M = 4.43$) compared to clinicians ($M = 2.35$). These results indicate the possibility that patient self-reports may overestimate therapist use of prescribed techniques but suggest that clinicians may be able to accurately assess their own utilization of BA techniques. Future studies will need to further evaluate the validity of these therapist and patient ratings for use as clinical support tools. Our comprehensive program of qualitative and quantitative approaches to the development of the BA fidelity assessments, indicate that these BA fidelity assessments have the potential to be important tools to support clinician training and research on BA in low resource settings.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10488-022-01219-w.

Declarations Research reported in this publication was supported by the National Institute of Mental Health (grants R21-MH116362 and R34-MH108818). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health. All participants provided written informed consent and the study was approved by the University of Pennsylvania Institutional Review Board committee #8. This manuscript follows the guidelines of a mixed methods approach. Data from this study is available upon reasonable request from the corresponding author. Mary Beth Connolly Gibbons, Jena Fisher, and Paul Crits-Christoph contributed to the study conception and design. Data analyses and preparation were performed by Robert Gallop. The first draft of the manuscript was written by Mary Beth Connolly Gibbons and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript. The authors report no competing interests.

References

Aarons, G. A., Hurlburt, M., & Horwitz, S. M. (2011). Advancing a conceptual model of evidence-based practice implementation in public service sectors. *Administration and Policy in Mental Health and Mental Health Services Research*, *38*(1), 4–23. https://doi.org/10.1007/s10488-010-0327-7

Bearsley-Smith, C., Sellick, K., Chesters, J., & Francis, K. (2008). Treatment content in child and adolescent mental health services: Development of the treatment recording sheet. *Administration and Policy in Mental Health and Mental Health Services Research*, *35*(5), 423–435. https://doi.org/10.1007/s10488-008-0184-9

Caron, E. B., & Dozier, M. (2022). Self-Coding of Fidelity as a Potential Active Ingredient of Consultation to Improve Clinicians' Fidelity. *Administration and Policy in Mental Health and Mental Health Services Research*, *49*(2), 237–254. https://doi.org/10.1007/s10488-021-01160-4

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. https://doi.org/10.1080/10503309812331332407

Carvalho, J. P., & Hopko, D. R. (2011). Behavioral theory of depression: Reinforcement as a mediating variable between avoidance and depression. *Journal of Behavior Therapy and Experimental Psychiatry*, *42*(2), 154–162. https://doi.org/10.1016/j.jbtep.2010.10.001

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, D., & Hays, R. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*, *63*(11), 1179–1194. https://doi.org/10.1016/j.jclinepi.2010.04.011

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., Ader, D., Fries, J. F., Bruce, B., & Rose, M. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Medical Care*, *45*(5 Suppl 1), S3–S11. https://doi.org/10.1097/01.mlr.0000258615.42478.55

Chapman, J. E., McCart, M. R., Letourneau, E. J., & Sheidow, A. J. (2013). Comparison of youth, caregiver, therapist, trained, and treatment expert raters of therapist adherence to a substance abuse treatment protocol. *Journal of Consulting and Clinical Psychology*, *81*(4), 674–680. https://doi.org/10.1037/a0033021

Christopher, M. S., Jacob, K. L., Neuhaus, E. C., Neary, T. J., & Fiola, L. A. (2009). Cognitive and behavioral changes related to symptom improvement among patients with a mood disorder receiving intensive cognitive–behavioral therapy. *Journal of Psychiatric Practice*, *15*(2), 95–102. https://doi.org/10.1097/01.pra.0000348362.11548.5f

Creed, T. A., Stirman, S. W., Evans, A. C., & Beck, A. T. (2014). A model for implementation of cognitive therapy in community mental health: The Beck Initiative. *Behavior Therapist*, *37*(3), 56–64

Crits-Christoph, P., Goldstein, E., King, C., Jordan, M., Thompson, D., Fisher, J., & Connolly Gibbons, M. B. (2021). A feasibility study of behavioral activation for major depressive disorder in a community mental health setting. *Behavior Therapy*, *52*(1), 39–52. https://doi.org/10.1016/j.beth.2020.01.008

Connolly Gibbons, M. B., Kurtz, J. E., Thompson, D. L., Mack, R. A., Lee, J. K., Rothbard, A., Eisen, S. V., Gallop, R., & Crits-Christoph, P. (2015). The effectiveness of clinician feedback in the treatment of depression in the community mental health system. *Journal of Consulting and Clinical Psychology*, *83*(4), 748–759. https://doi.org/10.1037/a0039302

Connolly Gibbons, M. B., Rothbard, A., Farris, K. D., Stirman, S. W., Thompson, S. M., Scott, K., Heintz, L. E., Gallop, R., & Crits-Christoph, P. (2011). Changes in psychotherapy utilization among consumers of services for major depressive disorder in the community mental health system. *Administration and Policy in Mental Health and Mental Health Services Research*, *38*(6), 495–503. https://doi.org/10.1007/s10488-011-0336-1

DeWalt, D. A., Rothrock, N., Yount, S., & Stone, A. A. (2007). Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care*, *45*(5 Suppl 1), S12–S21. https://doi.org/10.1097/01.mlr.0000254567.79743.e2

Dimidjian, S., Goodman, S. H., Sherwood, N. E., Simon, G. E., Ludman, E., Gallop, R., Welch, S. S., Boggs, J. M., Metcalf, C. A., Hubley, S., Powers, J. D., & Beck, A. (2017). A pragmatic randomized clinical trial of behavioral activation for depressed pregnant women. *Journal of Consulting and Clinical Psychology*, *85*(1), 26–36. https://doi.org/10.1037/ccp0000151

Dimidjian, S., Hubley, S., Martell, C., Herman, R., & Dobson, K. (2012). *Quality of behavioral activation scale*. Boulder, Colorado: Unpublished manuscript, University of Colorado Boulder

Ekers, D., Richards, D., McMillan, D., Bland, J. M., & Gilbody, S. (2011). Behavioural activation delivered by the non-specialist: Phase II randomised controlled trial. *The British Journal of Psychiatry*, *198*(1), 66–72. https://doi.org/10.1192/bjp.bp.110.079111

Ekers, D., Webster, L., Van Straten, A., Cuijpers, P., Richards, D., & Gilbody, S. (2014). Behavioural activation for depression; An update of meta-analysis of effectiveness and sub group analysis. *PLOS ONE*, *9*(6), e100100. https://doi.org/10.1371/journal.pone.0100100

Embretson, S. E., & Reise, S. P. (2000). *Multivariate Applications Books Series. Item response theory for psychologists*. Lawrence Erlbaum Associates Publishers

Gawrysiak, M., Nicholas, C., & Hopko, D. R. (2009). Behavioral activation for moderately depressed university students: Randomized controlled trial. *Journal of Counseling Psychology*, *56*(3), 468–475. https://doi.org/10.1037/a0016383

Gibbons, M. B. C., Thompson, S. M., Scott, K., Schauble, L. A., Mooney, T., Thompson, D., Green, P., MacArthur, M. J., & Crits-Christoph, P. (2012). Supportive-expressive dynamic psychotherapy in the community mental health system: a pilot effectiveness trial for the treatment of depression. *Psychotherapy*, *49*(3), 303–316. https://doi.org/10.1037/a0027694

Gitlin, L. N., Harris, L. F., McCoy, M. C., Chernett, N. L., Pizzi, L. T., Jutkowitz, E., Hess, E., & Hauck, W. W. (2013). A home-based intervention to reduce depressive symptoms and improve quality of life in older African Americans: A randomized trial. *Annals of Internal Medicine*, *159*(4), 243–252. https://doi.org/10.7326/0003-4819-159-4-201308200-00005

Hepner, K. A., Greenwood, G. L., Azocar, F., Miranda, J., & Burnam, M. A. (2010). Usual care psychotherapy for depression in a large managed behavioral health organization. *Administration and Policy in Mental Health and Mental Health Services Research*, *37*(3), 270–278. https://doi.org/10.1007/s10488-009-0247-6

Hill, C. E., O'Grady, K. E., & Elkin, I. (1992). Applying the Collaborative Study Psychotherapy Rating Scale to rate therapist adherence in cognitive-behavior therapy, interpersonal therapy, and clinical management. *Journal of Consulting and Clinical Psychology*, *60*(1), 73–79. https://doi.org/10.1037/0022-006X.60.1.73

Hogue, A., Dauber, S., & Henderson, C. E. (2014). Therapist self-report of evidence-based practices in usual care for adolescent behavior problems: Factor and construct validity. *Administration and Policy in Mental Health and Mental Health Services*, *41*(1), 126–139. https://doi.org/10.1007/s10488-012-0442-8

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 and Mental Health Services*, *42*(2), 229–243. https://doi.org/10.1007/s10488-014-0548-2

Hogue, A., Ozechowski, T. J., Robbins, M. S., & Waldron, H. B. (2013). Making fidelity an intramural game: Localizing quality assurance procedures to promote sustainability of evidence-based practices in usual care. *Clinical Psychology: Science and Practice*, *20*(1), 60–77. https://doi.org/10.1111/cpsp.12023

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 and Mental Health Services*, *37*(3), 230–244. https://doi.org/10.1007/s10488-009-0251-x

Kanter, J. W., Mulick, P. S., Busch, A. M., Berlin, K. S., & Martell, C. R. (2007). The Behavioral Activation for Depression Scale (BADS): Psychometric properties and factor structure. *Journal of Psychopathology and Behavioral Assessment*, *29*(3), 191–202. https://doi.org/10.1007/s10862-006-9038-5

Kanter, J. W., Rusch, L. C., Busch, A. M., & Sedivy, S. K. (2009). Validation of the Behavioral Activation for Depression Scale (BADS) in a community sample with elevated depressive symptoms. *Journal of Psychopathology and Behavioral Assessment*, *31*(1), 36–42. https://doi.org/10.1007/s10862-008-9088-y

Kanter, J. W., Santiago-Rivera, A. L., Santos, M. M., Nagy, G., López, M., Hurtado, G. D., & West, P. (2015). A randomized hybrid efficacy and effectiveness trial of behavioral activation for Latinos with depression. *Behavior Therapy*, *46*(2), 177–192. https://doi.org/10.1016/j.beth.2014.09.011

Karlin, B. E., Ruzek, J. I., Chard, K. M., Eftekhari, A., Monson, C. M., Hembree, E. A., Resick, A. P., & Foa, E. B. (2010). Dissemination of evidence-based psychological treatments for posttraumatic stress disorder in the Veterans Health Administration. *Journal of Traumatic Stress*, *23*(6), 663–673. https://doi.org/10.1002/jts.20588

Kelley, S. D., de Andrade, A. R. V., Sheffer, E., & Bickman, L. (2010). Exploring the black box: Measuring youth treatment process and progress in usual care. *Administration and Policy in Mental Health and Mental Health Services*, *37*(3), 287–300. https://doi.org/10.1007/s10488-010-0298-8

Manos, R. C., Kanter, J. W., & Busch, A. M. (2010). A critical review of assessment strategies to measure the behavioral activation model of depression. *Clinical Psychology Review*, *30*(5), 547–561. https://doi.org/10.1016/j.cpr.2010.03.008

Martell, C. R., & Addis, M. E. (2004). *Overcoming depression one step at a time: The new behavioral activation approach to getting your life back*. New Harbinger Publications

Martell, C. R., Dimidjian, S., & Herman-Dunn, R. (2013). *Behavioral activation for depression: A clinician's guide*. Guilford Press

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*(2), 181–193. https://doi.org/10.1080/10503300802688460

McManus, F., Rakovshik, S., Kennerley, H., Fennell, M., & Westbrook, D. (2012). An investigation of the accuracy of therapists' self-assessment of cognitive-behaviour therapy skills. *British Journal of Clinical Psychology*, *51*(3), 292–306. https://doi.org/10.1111/j.2044-8260.2011.02028.x

Miller, W. R., Yahne, C. E., Moyers, T. B., Martinez, J., & Pirritano, M. (2004). A randomized trial of methods to help clinicians learn motivational interviewing. *Journal of Consulting and Clinical Psychology*, *72*(6), 1050–1062. https://doi.org/10.1037/0022-006X.72.6.1050

Puspitasari, A. J., Kanter, J. W., Busch, A., Leonard, R., Dunsiger, S., Cahill, S., Martell, C., & Koerner, K. (2017). A randomized trial of an online, modular, active learning training program for behavioral activation for depression. *Journal of Consulting and Clinical Psychology*, *85*(8), 814–825. https://doi.org/10.1037/ccp0000223

Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., Markowitz, C. J., & Keller, M. B. (2003). The 16-item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry*, *54*(5), 573–583. https://doi.org/10.1016/s0006-3223(02)01866-8

Ryba, M. M., Lejuez, C. W., & Hopko, D. R. (2014). Behavioral activation for depressed breast cancer patients: The impact of therapeutic compliance and quantity of activities completed on symptom reduction. *Journal of Consulting and Clinical Psychology*, *82*(2), 325–335. https://doi.org/10.1037/a0035363

Schoenwald, S. K. (2011). It's a bird, it's a plane, it's… fidelity measurement in the real world. *Clinical Psychology: Science and Practice*, *18*(2), 142–147. https://doi.org/10.1111/j.1468-2850.2011.01245.x

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 and Mental Health Services*, *38*(1), 32–43. https://doi.org/10.1007/s10488-010-0321-0

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Southam-Gerow, M. A., & McLeod, B. D. (2013). Advances in applying treatment integrity research for dissemination and implementation science: Introduction to special issue. *Clinical Psychology: Science and Practice*, *20*(1), 1–13. https://doi.org/10.1111/cpsp.12019

Stein, A. T., Carl, E., Cuijpers, P., Karyotaki, E., & Smits, J. A. J. (2021). Looking beyond depression: a meta-analysis of the effect of behavioral activation on depression, anxiety, and activation. *Psychological Medicine*, *51*(9), 1491–1504. https://doi.org/10.1017/S0033291720000239

Weersing, V. R., Weisz, J. R., & Donenberg, G. R. (2002). Development of the therapy procedures checklist: A therapist-report measure of technique use in child and adolescent treatment. *Journal of Clinical Child and Adolescent Psychology*, *31*(2), 168–180. https://doi.org/10.1207/S15374424JCCP3102_03

Wells, R., Morrissey, J. P., Lee, I. H., & Radford, A. (2010). Trends in behavioral health care service provision by community health centers, 1998–2007. *Psychiatric Services*, *61*(8), 759–764. https://doi.org/10.1176/ps.2010.61.8.759