

UC San Diego

UC San Diego Previously Published Works

Title

Therapist-Observer Concordance in Ratings of EBP Strategy Delivery: Challenges and Targeted Directions in Pursuing Pragmatic Measurement in Children's Mental Health Services

Permalink

<https://escholarship.org/uc/item/7bz398d0>

Journal

Administration and Policy in Mental Health and Mental Health Services Research, 48(1)

ISSN

0894-587X

Authors

Brookman-Frazee, Lauren
Stadnick, Nicole A
Lind, Teresa
[et al.](#)

Publication Date

2021


DOI

10.1007/s10488-020-01054-x

Peer reviewed



Therapist-Observer Concordance in Ratings of EBP Strategy Delivery: Challenges and Targeted Directions in Pursuing Pragmatic Measurement in Children's Mental Health Services

Lauren Brookman-Frazee^{1,2,3} · Nicole A. Stadnick^{1,2,3}  · Teresa Lind^{1,2} · Scott Roesch⁴ · Laura Terrones^{1,2} · Miya L. Barnett⁵ · Jennifer Regan⁶ · Catherine A. Kennedy^{1,2} · Ann F. Garland⁷ · Anna S. Lau⁸

© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Pragmatic measures of therapist delivery of evidence-based practice (EBP) are critical to assessing the impact of large-scale, multiple EBP implementation efforts. As an initial step in the development of pragmatic measurement, the current study examined the concordance between therapist and observer ratings of items assessing delivery of EBP strategies considered essential for common child EBP targets. Possible EBP-, session-, and therapist-levels factors associated with concordance were also explored. Therapists and independent observers rated the extensiveness of therapist ($n = 103$) EBP strategy delivery in 680 community psychotherapy sessions in which six EBPs were used. Concordance between therapist- and observer-report of the extensiveness of therapist EBP strategy use was at least fair ($ICC \geq .40$) for approximately half of the items. Greater therapist-observer concordance was observed in sessions where a structured EBP was delivered and in sessions where therapists reported being able to carry out planned activities. Findings highlighted conditions that may improve or hinder therapists' ability to report on their own EBP strategy delivery in a way that is consistent with independent observers. These results can help inform the development of pragmatic therapist-report measures of EBP strategy delivery and implementation efforts more broadly.

Keywords Evidence-based practice implementation · Pragmatic measures · Community mental health

System-driven initiatives to implement multiple evidence-based practices (EBPs) are increasingly being used to

improve the quality and effectiveness of child mental health treatment in community settings (Lau and Brookman-Frazee 2016). Critical to understanding the impacts of these efforts is the ability to measure what therapists are actually delivering and specifically, the extent to which they are delivering EBP strategies. This measurement is essential to assess the effectiveness of community EBP implementation initiatives and to support quality assurance procedures (Kelley

Data described in this manuscript have been presented at scientific conferences.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10488-020-01054-x>) contains supplementary material, which is available to authorized users.

✉ Nicole A. Stadnick
nstadnic@health.ucsd.edu

¹ Department of Psychiatry, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

² Child and Adolescent Services Research Center, San Diego, CA, USA

³ University of California San Diego Dissemination and Implementation Science Center, La Jolla, CA, USA

⁴ Department of Psychology, San Diego State University, San Diego, CA, USA

⁵ Department of Counseling, Clinical, and School Psychology, University of California, Santa Barbara, Santa Barbara, CA, USA

⁶ Los Angeles County Department of Mental Health, Los Angeles, CA, USA

⁷ Department of Counseling & Marital and Family Therapy, University of San Diego, San Diego, CA, USA

⁸ Department of Psychology, University of California, Los Angeles, Los Angeles, CA, USA

et al. 2010; Schoenwald et al. 2011). Furthermore, accurate assessment of community-based practice is needed to test relationships between therapist EBP strategy delivery and child outcomes. Given the many competing demands in community mental health contexts, it is imperative that measurement of therapist EBP delivery be pragmatic, meaning that it is low burden and feasible for use in community practice settings, addresses stakeholder concerns, and is actionable and useful for advancing efforts in training, supervision, quality monitoring, and/or quality improvement (Glasgow and Riley 2013; National Institute of Mental Health 2017).

The development of community practice measures has been a challenge. Currently available methods fall into three primary categories: (1) observational coding systems, (2) client report, and (3) therapist report (Orimoto et al. 2012). Observational assessments are considered the gold standard in measuring treatment fidelity in controlled trials as they are thought to provide the most objective and thorough measurement of therapist behavior. However, this method is highly time-intensive, and unlikely to be feasible in routine care (Schoenwald et al. 2011). Alternatively, client and caregiver reports require considerably less time and resources, but present challenges stemming from clients' naïveté to the nature and vocabulary pertaining to treatment strategies delivered (Regan 2014). The third option, therapist-report of their own EBP delivery, is promising, though establishing reliability between therapist and observational ratings has proved challenging (Hurlburt et al. 2010). The current study reports on the first phase of a multi-phase process of the development of a pragmatic therapist-report measure of EBP strategy delivery in the context of a system-driven implementation of multiple EBPs in routine care for children and adolescents. As such, it describes therapist and observer ratings on a set of preliminary items assessing EBP strategy delivery, and explores concordance between reporters (observers and therapist self-reporters). Although we do not report on a finalized therapist report measure, the findings offer guidance for next steps in measure development by identifying factors that may influence the reliability of therapist report of EBP strategy delivery in comparison to independent observer ratings.

Challenges of Pragmatic Measurement of EBP Strategy Delivery: Therapist Self-Report

Despite concerns about the reliability of therapist self-report of EBP delivery (Hurlburt et al. 2010), therapist report has been recommended as a pragmatic and feasible method of assessing intervention delivery (Hogue et al. 2015). Therapist report measures offer several advantages, including

alignment with integrity measures used in the context of EBP training and implementation (Borntreger et al. 2015; Hogue et al. 2015), promotion of self-reflection upon EBP delivery that can have downstream benefits for implementation, a basis for performance feedback from supervisors, and incorporation into evaluation of therapist and agency performance objectives (Bearsley-Smith et al. 2008). Furthermore, community therapist self-reports of their EBP delivery have been shown to predict changes in client-level mental health outcomes (Dopp et al. 2017). In addition, such measures could also be used in quality assurance, in which measurement and feedback on factors such as treatment progress (e.g., youth outcomes) and treatment processes (e.g., treatment activities) can be used to enhance clinical decision-making, improve accountability, and drive program planning (Chorpita et al. 2008; Garland et al. 2010a; Kelley and Bickman 2009; McLeod et al. 2013). Thus, therapists' reports of their EBP delivery offer a feasible approach with a range of potential benefits in scale-up efforts. Unfortunately, there is evidence of limited agreement between therapist self-report and observer ratings of therapist EBP delivery (Hurlburt et al. 2010). This limited agreement between therapist self-report and observer ratings has been noted with single EBPs, as well as with measures designed to assess multiple EBP delivery (Borntreger et al. 2015). Thus, research is needed to understand when therapist report is most aligned with independent observers to inform development of pragmatic measurement approaches.

Challenges of Pragmatic Measurement of EBP Strategy Delivery: Multiple EBP Context

Broad system-driven implementation efforts often include simultaneous implementation of multiple EBPs to address the range of needs of clients presenting for treatment (Lau and Brookman-Frazee 2016; Nakamura et al. 2011). This context provides a challenge to characterizing EBP delivery (Beidas et al. 2013; Chorpita et al. 2011). Although EBP-specific fidelity measures are essential for focused examination of EBP delivery, within multiple EBP implementation contexts, therapist delivery of essential strategies related to EBPs may be an optimal unit of measurement (Chorpita and Daleiden 2009). This approach is consistent with the "common-elements" approach to identifying treatment components shared across many effective interventions for a given mental health target (Boustani et al. 2017; Garland et al. 2008b), and has been used to characterize EBP strategy use in usual care psychotherapy for children (Garland et al. 2010b). Similarly, the evidence-informed model of care, Managing and Adapting Practice (MAP), and the modular EBP, MATCH, have leveraged therapist report of therapeutic

practice delivery within clinical dashboards, via consultation records, and within a Monthly Treatment and Progress Summary (MTPS) (Chorpita and Daleiden 2009). Although, reported treatment components via consultation records have been validated against observer ratings within an effectiveness trial (Ward et al. 2012), validation of therapist reports on the MTPS is thus far limited to 9 of 55 practice elements (Borntrager et al. 2013).

Developing a Pragmatic Therapist-Report Measure of Multiple EBP Strategy Delivery

In response to the need to for a pragmatic assessment of EBP delivery and building on the common elements approach, our group initiated a multi-phase study aimed to develop a pragmatic assessment of multiple EBP delivery. Specifically, our goal is to develop a therapist-report measure of EBP concordant care, referring to the extent to which a therapist delivers treatment strategies considered essential in EBP protocols for a specific mental health target or family of treatments (e.g., trauma). The end goal is to produce a therapist-report measure assessing the extensiveness of a range of individual EBP strategies likely observable in many treatment sessions and familiar to therapists trained in an EBP within the mental health problem focus (Cho et al. 2019; Garland et al. 2010b; Orimoto et al. 2012). This approach is intended to complement the use of EBP-specific fidelity measures that remain essential for initial EBP training and evaluation of individual EBPs. This study reports findings from the first steps in examining therapist-observer concordance for a preliminary set of items assessing EBP strategy delivery on the EBP Concordance Care Assessment-alpha version (ECCA- α). The ECCA- α is not a finalized measure ready for use in community practice, but initial results provide information to derive needed improvements to generate a pragmatic and psychometrically strong measure of EBP strategy delivery.

Examining Therapist and Observer Concordance

There are very few studies that have examined therapist-observer concordance of treatment strategy delivery within the context of community EBP implementation. The studies that have been conducted have reported variation in concordance. Some research, for example, suggests that therapist reports diverge from objective observer ratings (Brosan et al. 2008; Herschell et al. 2019; Hurlburt et al. 2010; Miller et al. 2004). Across studies of child and adult EBPs, therapists tend to report higher rates of EBP strategy use than observers (Carroll et al. 1998; Decker and Martino

2013; Herschell et al. 2019; Hogue et al. 2015; Hurlburt et al. 2010). Other studies have reported adequate concordance between therapist and observer ratings (Hogue et al. 2014; Martino et al. 2009; Schoenwald et al. 2011). Hogue et al. (2014), for example, found that therapists were able to reliably report on the amount of time devoted in session to specific EBP treatment targets and did not over-report compared to observers within a family-based preventive intervention for adolescents. It is unknown how therapists and observer ratings correspond within the context of multiple EBP implementation.

Predicting Observer-Therapist Concordance

In addition to characterizing observer-therapist concordance, it is also critical to identify factors that may influence therapist and observer alignment in rating therapist EBP strategy use. Though research in this area is limited, some data suggest that the type of EBP, therapist level of training, and therapist reflective ability may influence observer-therapist reliability (Hogue et al. 2015; Loades and Myles 2016). In comparing types of EBPs, Hogue et al. (2015) found that overall observer-therapist reliability for therapist EBP strategy use was adequate for family therapy sessions ($ICC = 0.66$), but very low for motivational interviewing and CBT sessions ($ICC = 0.06$). In addition, trainee therapists who had higher levels of reflective ability tended to agree more closely with observer ratings of their strategy use, and trainee therapists did not tend to over-estimate their strategy use in the ways observed among more experienced therapists (Loades and Myles 2016). Also related to training, Hogue et al. (2015) noted that community therapists untrained in family therapy tended not to inflate their self-ratings of family therapy strategies relative to observers. While it is likely that therapist and observers will have varying levels of agreement on therapist EBP delivery given their different access to information about a given session and client, understanding when therapists and observers agree most is essential information for developing pragmatic therapist-report measures of EBP delivery.

Current Study

This initial study was conducted within the context of children's mental health services provided through the Los Angeles County's Department of Mental Health (LACDMH) during the large-scale implementation of multiple EBPs. The aims of the current study were to (1) describe patterns of observer and therapist ratings on items assessing therapist EBP strategy delivery; (2) characterize concordance between therapist and observer ratings of these items; and (3) identify

factors associated with overall therapist-observer concordance. Findings from this initial study will provide targeted direction in the further development of a pragmatic EBP measure and can inform the development of similar measures for other service contexts.

Method

Recruitment and Procedures

Data for the current study were drawn from the “in-depth” sample of the Knowledge Exchange on Evidence-based Practice Sustainment (“4KEEPS”) study (Lau and Brookman-Frazer 2016). This study was conducted in 24 programs within 14 agencies contracted with LACDMH to deliver EBPs to children and adolescents. Programs were eligible if they were contracted to deliver at least one of the six EBPs of interest and had at least five therapists delivering them. These (selected by LACDMH for large scale training) included Cognitive Behavioral Intervention for Trauma in Schools (CBITS), Child-Parent Psychotherapy (CPP), Seeking Safety (SS), Trauma-Focused Cognitive Behavior Therapy (TF-CBT), Triple P—Positive Parenting Program (Triple P), as well as Managing and Adapting Practice (MAP).¹ Therapists were eligible if they were: (1) employed as a staff or trainee therapist in a participating program site, (2) trained in one of the six EBPs of interest, and (3) currently delivering one of these EBPs to at least one client.

Therapists were recruited for the study at staff meetings held at their program site, and participated in an initial survey and interview. Therapists were instructed to submit audio recordings of up to three sessions per client within a six week period, for up to three clients. For each recorded session, therapists were asked to complete the therapist report version of the ECCA- α (see description below) as soon as possible following the session. Sessions were included in the data analysis if the ECCA was submitted within one week of the session.

¹ We use the term “EBPs” throughout, inclusive of MAP, for ease of communication. We acknowledge that MAP differs from the other EBPs as it is an evidence management system, with a direct service component that coordinates multiple evidence sources, employs practices components drawn from over 700 evidence based treatments, and offers structured process management supports to guide clinical care for anxiety, trauma, depression, conduct problems in children 0–21 years old. In the LACDMH system, the MAP direct service model complements a diverse array of EBPs by allowing providers to select, review, adapt, or construct promising treatments as needed to match particular child characteristics based on the latest scientific findings (see Chorrita and Daleiden 2018).

Therapists received a \$20 incentive for completing the initial survey, a \$10 incentive for each session recording, and an additional \$10 for completing the questionnaires associated with the session. They were also permitted to keep the iPod used as a recording device following study completion as incentive for submission of data for at least 6 sessions. All procedures were approved by the Institutional Review Boards at the University of California San Diego and the University of California Los Angeles and by the Los Angeles County Department of Mental Health.

Participants

One hundred and three therapists were included in this study, with data from 680 session audio recordings for 273 child clients. On average, therapists submitted data for 2.65 children each (*SD* 0.64), with an average of 2.49 sessions submitted per child (*SD* 0.73). Additional details regarding session, child, and therapist characteristics are in Table 1. Session audio recordings were included in the study if they met the following eligibility criteria: at least 15 min in length, intelligible audio quality, and linked with a corresponding therapist-report ECCA- α completed within seven days of the session.

Measures

Session-Level Measures

Session Participants For each session submitted, therapists indicated who (caregiver, youth, or both) was present in each session.

Session EBP Focus Therapists indicated which of 6 EBPs was delivered in the session—CBITS, CPP, MAP (specifying MAP Anxiety, MAP Conduct, MAP Depression, MAP Trauma), Triple P, TF-CBT, or SS. Practices were grouped based on their primary mental health treatment target into: Trauma (CBITS, CPP, MAP Trauma, TF-CBT, Seeking Safety), Conduct (Triple P, MAP Conduct), Anxiety (MAP Anxiety), and Depression (MAP Depression).

EBP Characteristics Based on a previous study in which EBPs were classified based on their intervention and implementation characteristics (Barnett et al. 2017), two mutually exclusive dichotomous variables were included to represent: (1) whether the EBP had prescribed session content and order, and (2) whether the implementation requirements for the EBP included ongoing consultation.

Ability to Carry Out Intended Session Activities Therapists were asked to rate the extent to which they were able

Table 1 Descriptives of session, child, and therapist characteristics

Characteristics	Session (<i>n</i> = 680)	Child (<i>n</i> = 273)	Therapist (<i>n</i> = 103)
<i>Session EBP, no. (%)</i>			
CPP	49 (7.2)	–	–
CBITS	0 (0)	–	–
MAP	357 (52.5)	–	–
MAP—anxiety	111 (16.3)	–	–
MAP—conduct	153 (22.5)	–	–
MAP—depression	83 (12.2)	–	–
MAP—trauma	10 (1.5)	–	–
SS	27 (4.0)	–	–
TF-CBT	209 (30.7)	–	–
Triple P	38 (5.6)	–	–
Caregiver present in session, no. (%)	285 (41.9)	–	–
Language other than English, no. (%)	116 (17.1)	–	63 (61.2)
Female gender, no. (%)	–	138 (50.5)	91 (88.3)
<i>Race/ethnicity, no. (%)</i>			
Non-Hispanic White	–	13 (4.8)	22 (21.4)
Hispanic	–	193 (70.7)	57 (55.3)
Other ethnic minority	–	67 (24.5)	24 (23.3)
Age, <i>M</i> (<i>SD</i>); range	–	9.8 (3.9); 1–18 years	–
<i>Primary diagnosis, no. (%)</i>			
Anxiety	–	50 (18.3)	–
Attention or hyperactivity problems	–	7 (2.6)	–
Mood	–	55 (20.1)	–
Trauma	–	75 (27.5)	–
Disruptive behavior or conduct	–	79 (28.9)	–
Other	–	7 (2.6)	–
<i>Education, no. (%)</i>			
Less than Master's Degree	–	–	4 (3.9)
Master's Degree	–	–	88 (85.4)
Doctoral Degree	–	–	11 (10.7)
Therapist Licensed, no. (%)	–	–	20 (19.4)
Years of professional experience, <i>M</i> (<i>SD</i>)	–	–	4.4 (4.4)
Number of EBPs trained in, <i>M</i> (<i>SD</i>)	–	–	2.3 (.9)

There were no sessions of CBITS included in this study, as this intervention was not widely implemented throughout the county

EBP evidence-based practice, *CPP* Child-Parent Psychotherapy, *CBITS* Cognitive Behavioral Intervention for Trauma in Schools, *MAP* Managing and Adapting Practice, *SS* Seeking Safety, *TF-CBT* Trauma-Focused Cognitive Behavioral Therapy, *Triple P* Positive Parenting Program

to carry out session activities as intended in each session on a six-point Likert scale (1 = “not at all” to 6 = “fully”).

Client-Level Measures

Client Characteristics Therapists reported basic client demographic information including age, gender, and race/ethnicity of the child client.

Therapist-Level Measures

Therapist Demographic, Professional, and Practice Characteristics Therapists completed the Therapist Background Questionnaire (Brookman-Frazer et al. 2012) concerning personal and professional characteristics, including the number of EBPs in which they were trained and the number of direct services hours provided per week.

Therapist Emotional Exhaustion Therapists responded to 5 items from the Emotional Exhaustion subscale of the Organizational Social Context Questionnaire (OSC; Glisson et al. 2008). Therapists rated their perceptions of stressful climates characterized by factors such as workload (e.g., “I feel used up at the end of the day”) and work-related emotional exhaustion (e.g., “I feel emotionally drained from my work”). Responses were on 7-point Likert Scale (0 = strongly disagree, 6 = strongly agree) with higher scores representing more emotional exhaustion. The current sample demonstrated good internal consistency of $\alpha=0.81$ using Cronbach’s alpha.

Therapist General Attitudes Towards EBPs Therapists’ general attitudes towards the adoption of EBPs was measured with the Evidence-Based Practice Attitudes Scale (EBPAS; Aarons 2004). The current study included two subscales: openness and divergence, each of which consisted of four items. The openness subscale assessed the therapist’s openness to trying new interventions and willingness to use EBPs (e.g., “I like to use new types of therapy/interventions to help my clients”). The divergence subscale assessed the perception that EBPs were not as useful as clinical experience (e.g., “Research based treatments/interventions are not clinically useful”). Therapists rated each item on a five-point Likert scale (0 = *not at all*, 4 = *very great extent*). In the current sample, the Cronbach’s alpha indicated that the internal consistency was acceptable for the openness scale ($\alpha=0.79$) and for the divergence scale ($\alpha=0.71$).

Therapist Perceptions of Specific EBPs Therapists completed an adapted version of the Perceived Characteristics of Intervention Scale (PCIS; Cook et al. 2015) for any EBP they had ever received training or delivered. This version of the PCIS included 8 items related to relative advantage (e.g., “[The practice] is more effective than other therapies I have used”), compatibility (e.g., “[The practice] is aligned with my clinical judgment”), complexity (e.g., “[The practice] is easy to use”), and potential for reinvention (e.g., “[The practice] can be adapted to meet the needs of my client”). Therapist rated their agreement with each item on a 5-point Likert scale (1 = not at all, 5 = a very great extent). Internal consistencies of the 8-item scale for all practices in the current sample were strong, with alphas ranging from $\alpha=0.93$ to $\alpha=0.97$. This study utilized the PCIS score regarding the EBP delivered in the session.

Therapist EBP Delivery Self-Efficacy Two items assessed therapist perceptions of confidence in using a specific EBP. The two items were: “I am well prepared to deliver [Practice] even with challenging clients,” and “I am confident in my ability to implement [Practice].” Therapists answered these items about each EBP that were trained in. Each item

was rated on a 5-point Likert scale (1 = not at all, 5 = a very great extent). A sum of both items was used as a composite score (scores ranged from 2–10). Pearson correlation coefficients ranged from 0.75 to 1.0 across EBPs. The current study used the mean self-efficacy composite score for the EBP being delivered in that session.

Initial Items in the Evidence-Based Practice Concordant Care Assessment (ECCA- α): Therapist Report and Observer Report

The items included in the ECCA- α represented an initial effort to measure the extent to which therapists deliver a set of EBP strategies considered essential across evidence-based protocols for common child mental health targets (trauma, conduct problems, anxiety, depression). Items in the ECCA- α were derived from a practice expert survey and several previous measurement systems, including the Practice and Research: Advancing Collaboration Therapy Process Observational Coding System for Child Psychotherapy-Specific Therapy Process Scale (PRAC TPOCS-S; Garland et al. 2008b) and the Monthly Treatment and Progress Summary (MTPS; Child and Adolescent Mental Health Division 2003). The ECCA- α captured EBP strategies from six EBPs of interest: CPP, CBITS, TF-CBT, SS, MAP, and Triple P. See the Online Appendix which describes the development of these items, including sources.

The ECCA- α includes 32 items assessing the occurrence and extensiveness of therapist delivery of strategies considered essential for four common child mental health problem targets—trauma, conduct, anxiety, and depression. Items assess both *content* (24 items; e.g., time-out, exposure, trauma narrative, activity scheduling) and *techniques* (8 items) used to deliver content (e.g., agenda setting, homework assignment and review, modeling, role play). Consistent with the PRAC TPOCS-S (Garland et al. 2008a), content items were defined as “the substance or issue being addressed in the therapeutic intervention,” whereas technique items that were defined as “the active method or the way a therapist attempts to intervene with, or relate to, a client.” Specific content areas may be covered using different techniques and specific techniques may be used to cover many different content areas. Items were rated on a 7-point Likert scale reflecting occurrence and extent to which the strategy was used in a given session and ranged from 0 (not used) to 6 (used with great extent).

Ratings Consistent with the Garland and colleagues’ (2008) approach to characterizing strategy delivery using multiple metrics, the following item ratings are generated for each session from the 7-point Likert scale used by raters: (a) a dichotomous “Occurrence Rating” for each of the 32 items with 0 indicating “not used” (i.e., strategy rated as a 0) and

1 indicating that the strategy was used (i.e., strategy was rated between 1 and 6); (b) a continuous “Extensiveness Rating” for each of the 32 items ranging from 0 to 6 (i.e., the same as the raw Likert Rating Scale); and (c) a continuous “Extensiveness-When-Occurred Rating” for the subset of items rated as used (i.e., > 0 , with a possible score range of 1 to 6). Each of these types of item ratings were generated for all of the items for each session that was coded.

Problem Target Composite Scoring Four *Problem Target Composites* were created by calculating the mean of the Extensiveness ratings for a subset of items related to focus of the EBP delivered during a particular session (i.e., trauma, conduct, anxiety, or depression). Inclusion of items on each of the Problem Target Composites were determined based on practice expert ratings (see details in the Online Appendix). Individual items could be included on more than one Problem Target Composite (Trauma—15 items; Anxiety—14 items; Conduct—26; Depression—15). Composite scores ranged from 0–6 with higher composite scores indicating greater extensiveness of EBP strategy delivery for a given problem target of an individual session. See the Online Appendix for additional details.

Session ECCA- α Therapist Report The therapist-report ECCA- α was administered to therapists through an online survey which was completed within a week of a particular session. Therapists were instructed to indicate which strategies they had used in that specific session. For each EBP strategy, therapists were asked to rate their extensiveness of use of the strategy on a scale from 0 to 6, with 0 indicating “not at all,” 3 indicating “to a moderate extent, and 6 indicating “to a great extent.”

Session ECCA- α Observer Report The ECCA- α Observational Coding System Manual included the same items as the Therapist Report ECCA- α . These items included definitions of strategies that were the same in both the therapist and observer versions of the ECCA- α , although the ECCA- α Observational Coding System included additional examples. The structure of the ECCA- α Observational Coding System was adapted from the PRAC TPOCS-S (Garland et al. 2008a). Consistent with the therapist version, coders were instructed to rate the extensiveness of strategy use on a scale from 0 to 6, with 0 indicating “not at all,” 3 indicating “to a moderate extent,” and 6 indicating “to a great extent.” Coders were instructed to consider two related dimensions in rating the extensiveness of strategy use: (1) the thoroughness of the strategy use (including effort, detail, depth/intensity, and follow-through), and (2) the frequency of the strategy use (number of instances used during a session). Additional detail regarding coding instructions are found in the Online Appendix.

A total of 13 research staff (62% undergraduate, 38% post-baccalaureate) were trained in the ECCA- α Observational Coding System and coded audio recordings of each session. Coder training included manual review, didactic training sessions, and practice coding. Coders were considered reliable and ready to start independent coding when they reached achieved at least 80% agreement for each of the six or more criterion rated recordings. Coders were randomly assigned to each session and were kept naïve to the EBP being delivered and the problem target addressed. Twenty-six percent of the sessions were randomly selected for double coding for purposes of evaluating inter-rater reliability. A representative sample of sessions for each of the 13 coders was included in the sessions selected for double-coding. Sessions conducted in Spanish were coded by two of the reliable coders who were fluent in Spanish.

Observer inter-rater reliability was calculated using a one-way random effects $ICC_{(1,k)}$ model based on a mean-rating, absolute-agreement (Hallgren 2012; McGraw and Wong 1996). This model was considered appropriate because a variety of rater pairs were randomly sampled from the overall group of raters to double-code a subset of sessions (Hallgren 2012; McGraw and Wong 1996). Because the rater pairs were randomly sampled, the effect is random and generalizes to the larger population of raters (McGraw and Wong 1996). However, because calculations were based on different subjects being rated by different subsets of coders, the assignment was not a fully crossed design (Hallgren 2012). This design means that $ICCs$ may underestimate the true reliability, as it was not possible to assess and control for systematic bias between coders. Using the standards outlined by (Cicchetti 1994) classifying $ICCs$ below .40 as reflecting “poor” agreement, $ICCs$ from .40 to .59 reflecting “fair” agreement, $ICCs$ from .60 to .74 reflecting “good” agreement, and $ICCs$.75 and higher reflecting “excellent” agreement, inter-rater reliability between observer coders for the 32 items was acceptable (all ICC ’s of at least 0.40), and 30 of the items had inter-rater reliability estimates in the good to excellent range (ICC ’s of at least 0.60). The average $ICC_{(1,k)}$ for all items was 0.74 ($SD = 0.11$; range from 0.44 (Monitoring) to 0.92 (Trauma Narrative). Please see Table 3 in the Online Appendix for the $ICC_{(1,k)}$ values for each item.

Data Analytic Plan

Aim 1 sought to describe the patterns of occurrence and extensiveness of observer and therapist ratings of items assessing therapist EBP strategy delivery. Descriptive statistics were used to characterize the occurrence and extensiveness of individual strategies for each rater type: (a) Frequency of individual strategy occurrence (the percent of sessions in which Occurrence = 1); (b) Average *Extensiveness-when-occurred* scores for individual strategies (range

1–6); (c) Average *Extensiveness* scores for each item (range 0–6). In addition to individual items, descriptive statistics were used to examine the average Problem Target Composite scores.

Aim 2 sought to characterize the concordance between observer and therapist ratings of EBP strategy delivery regardless of the specific EBP delivered in a session. To examine agreement between therapists and observers for each item, one-way random effects intra-class correlation coefficient estimates ($ICC_{(1,k)}$) were calculated using IBM SPSS Statistics for Windows (Version 25.0) based on a mean-rating, absolute-agreement (Koo and Li 2016). Consistent with the criteria of Cicchetti (1994), the cutoff for fair agreement was set at $ICC \geq 0.4$. In addition to item-level $ICCs$, we evaluated therapist-observer concordance with interrater q -correlations, which provided an index of overall agreement in the item-by-item pattern of therapist and observer ratings across all the EBP strategies for a given composite (Lau et al. 2004; Youngstrom et al. 2000). Q -correlations were calculated by applying the formula for Pearson r correlations between therapist-observer pairs, indexing item by item agreement across the common item set (the items comprising the Problem Target Composite for a particular session; (Fung and Lau 2010; Youngstrom et al. 2000). Q -correlations are sensitive to the shape and dispersion of the profile of item scores, quantifying the agreement between therapist and observer ratings across all items. The larger the q -correlation, the stronger the level of overall convergence in therapist and observer ratings. A zero correlation, or one near zero, indicates that therapist and observer ratings do not covary in any predictable, linear fashion.

For Aim 3, predictors of observer-therapist concordance were identified. A multilevel model was run to examine the effects of session-, client-, and therapist-level predictors on observer-therapist agreement as measured by q -correlation. In predicting q -correlations, a significant positive predictor indicated that an increase in the predictor score was associated with an increase in the level of overall agreement between therapist and observer ratings as a pattern across all items (Youngstrom et al. 2000). Due to the nested nature of the data (sessions within clients within therapists within agencies), unconditional models were run to determine whether there was significant variance attributable to the child, therapist, and agency levels. A significant proportion of variance in concordance was attributable to the client level ($ICC = 0.29$), the therapist level ($ICC = 0.21$), and a small proportion of variance in concordance was attributed to the agency level ($ICC = 0.04$). To account for the nested structure of the data, analyses employed a four-level model with session observations (Level 1; $n = 680$), nested within clients (Level 2; $n = 273$), nested within therapists (Level 3; $n = 103$), nested within agencies (Level 4; $n = 14$). All

multilevel analyses were run using Stata Statistical Software (Special Edition Release 15).

Results

Aim 1: Describe Patterns of Occurrence and Extensiveness of Observer and Therapist Ratings of EBP Strategy Delivery

Refer to Table 2 for the average extensiveness ratings and rates of occurrence (i.e., whether the strategy was delivered in the session at any level of extensiveness) for therapist and observer reports of individual strategies. For techniques, therapists most frequently reported using the strategies of End of Session Positive (in 67% of sessions), Establishing/Reviewing Agenda or Treatment Goals (in 65% of sessions), and Delivering Positive Reinforcement and Rewards (in 59% of sessions). Observers most frequently endorsed use of the techniques Establishing/Reviewing Agenda or Treatment Goals (in 91% of sessions), Delivering Positive Reinforcement and Rewards (in 88% of sessions), and Psychoeducation (in 82% of sessions). For content, therapists most frequently endorsed using Praise (in 56% of sessions), Communication and Social Skills (in 54% of sessions), Monitoring (in 42% of sessions), and Social Skills (in 42% of sessions). The most frequently endorsed content strategies by observers were Affect Education (in 69% of sessions), Relaxation (38% of sessions), and Cognitive Restructuring (in 26% of sessions).

When a particular strategy was endorsed, therapists rated their extensiveness in the “moderate” range, on average ($M = 3.75$, $SD = 0.32$). In contrast, observers rated therapist extensiveness for techniques below “moderate” extensiveness ($M = 2.80$, $SD = 0.37$). This pattern was consistent in the therapist and observer ratings of extensiveness for content items, with therapists, on average, rating extensiveness as 3.45 ($SD = 0.33$), and observers, on average, rating extensiveness as 2.61 ($SD = 0.36$).

Aim 2: Characterize Concordance Between Observer and Therapist Ratings of EBP Strategy Delivery

To examine concordance between therapist and observer ratings on each ECCA- α item, one-way random effects intra-class correlation coefficient estimates ($ICC_{(1,k)}$) were calculated based on a mean-rating, absolute-agreement (Koo and Li 2016). See Table 2 for the $ICC_{(1,k)}$ estimates for each individual ECCA item. The average $ICC_{(1,k)}$ across all items was 0.35 ($SD = 0.23$) with a large range (-0.23 to 0.78). Fifty six percent of the items (14 of 32) had therapist-observer $ICCs \geq 0.4$, which we considered the minimum acceptable threshold

Table 2 Therapist and observer reports of EBP strategy delivery: occurrence and average extensiveness for individual strategies and concordance between raters

EBP strategy	Strategies indicated for EBP target ^a	Therapist-observer inter-rater reliability $ICC_{(1,k)}$ ^b Est. (95% CI)	Therapist report		Observer report	
			% ^c	<i>M</i> (<i>SD</i>) ^d	% ^c	<i>M</i> (<i>SD</i>) ^d
<i>Techniques related to organizing/structuring treatment</i>						
Establishing/reviewing agenda or treatment goals	T, C, A, M	.38 (.28, .47)	64.7	3.53 (1.46)	91.2	3.06 (1.43)
Psychoeducation	T, C, A, M	.41 (.32, .49)	59.7	3.52 (1.50)	82.1	3.14 (1.52)
Tracking/reviewing client's progress	T, C, A, M	.33 (.22, .43)	56.3	3.51 (1.47)	56.9	2.35 (1.42)
<i>Techniques related to skill building</i>						
Modeling	C, A	.23 (.11, .34)	56.3	3.52 (1.41)	51.0	2.62 (1.53)
Role play & practice	T, C, A, M	.49 (.41, .57)	46.6	3.70 (1.47)	57.1	3.26 (1.56)
Assigning/reviewing homework	T, C, A, M	.64 (.58, .69)	48.8	3.89 (1.50)	49.4	3.07 (1.47)
Delivering positive reinforcement & rewards	C	.35 (.25, .44)	59.3	3.91 (1.49)	87.6	2.61 (1.39)
<i>Techniques related to engaging client & family</i>						
End of session positive	C, A, M	-.23 (-.44, -.05)	67.1	4.44 (1.52)	28.3	2.32 (1.32)
<i>Content related to behavioral parent training</i>						
Stimulus/antecedent control	C	.01 (-.26, .21)	36.7	3.11 (1.51)	11.7	2.12 (1.19)
Praise	T, C, A	.43 (.28, .55)	55.9	3.78 (1.55)	23.5	2.94 (1.19)
Tangible rewards	C	.43 (.27, .54)	31.0	3.72 (1.53)	19.6	2.75 (1.54)
Ignoring/differential reinforcement of other behaviors	C	.62 (.52, .70)	27.4	3.38 (1.69)	12.8	2.72 (1.83)
Attending	C	-.01 (-.28, .20)	36.7	3.72 (1.63)	12.1	2.82 (1.22)
Natural & logical consequences	C	.43 (.33, .51)	22.1	3.43 (1.51)	14.7	2.47 (1.48)
Time out	C	.78 (.72, .83)	18.9	3.62 (1.60)	11.0	3.19 (1.60)
Commands	C	.38 (.21, .51)	29.9	3.40 (1.55)	14.6	3.12 (1.63)
Behavioral contracting	C	.22 (.09, .33)	15.0	3.58 (1.50)	6.2	2.83 (1.51)
<i>Content related to cognitive and behavioral skills</i>						
Monitoring	T, C, A, M	.07 (-.08, .20)	41.8	3.32 (1.51)	16.0	2.39 (1.47)
Problem solving skills	C, M	.34 (.24, .44)	41.8	3.34 (1.48)	20.1	2.54 (1.63)
Exposure	T, A	.13 (-.01, .25)	26.3	3.31 (1.50)	8.1	2.35 (1.25)
Cognitive restructuring	T, C, A, M	.53 (.46, .60)	33.2	3.41 (1.56)	25.6	2.97 (1.71)
Activity scheduling	M	.23 (.11, .34)	28.8	3.32 (1.43)	10.1	2.61 (1.56)
Education support/academics	C	.29 (.17, .39)	12.9	2.78 (1.50)	10.9	2.16 (1.26)
Maintenance/relapse prevention	T, C, A, M	.17 (.04, .29)	22.8	3.30 (1.50)	5.7	2.49 (1.37)
<i>Content related to relating to others</i>						
Communication and social skills	C, M	.21 (.09, .32)	54.1	3.40 (1.51)	27.4	2.39 (1.52)
Assertiveness training	M	.14 (.002, .26)	27.6	2.88 (1.53)	6.2	2.19 (1.45)
<i>Content related to emotion regulation</i>						
Relaxation	T, C, A, M	.73 (.69, .77)	40.9	3.72 (1.59)	38.2	2.95 (1.74)
Caregiver coping	C	.22 (.01, .38)	35.4	3.35 (1.55)	13.9	1.92 (1.16)
Affect education	T, C, A, M	.42 (.30, .52)	40.5	3.77 (1.63)	68.9	2.68 (1.51)
<i>Content related to trauma and safety</i>						
Common reactions to trauma	T	.56 (.49, .62)	26.3	3.31 (1.50)	14.4	3.04 (1.62)
Trauma narrative	T	.76 (.72, .79)	12.4	4.48 (1.85)	12.4	2.92 (1.61)
Safety skills	T	.40 (.30, .48)	15.1	3.31 (1.74)	12.9	2.00 (1.41)

Items with an acceptable level of agreement ($ICC_{(1,k)} \geq .40$) are indicated in bold

^aT = Trauma practices; C = Conduct practices; A = Anxiety practice; M = Mood practice

^b $ICC_{(1,k)}$ indicates therapist-observer inter-rater reliability. Therapist-observer inter-rater reliability was calculated with a one-way random effects intra-class correlation model, $ICC_{(1,k)}$, mean-rating and absolute-agreement

^c% is the percentage of sessions in which the respondent indicated that the therapist used the strategy

^d*M* reflects the average extensiveness of the strategy when it occurred (i.e., extensiveness rating between 1 and 6), and *SD* reflects the standard deviation

Table 3 EBP target composites: therapist and observer reports

EBP target composite ^a	Therapist report <i>M (SD)</i>	Observer rating <i>M (SD)</i>	<i>t</i>	Q-correlation (agreement score)				
				<i>M (SD)</i>	Range	Percentiles		
						25th	50th	75th
Trauma composite (<i>n</i> = 295)	1.55 (0.96)	1.18 (0.64)	- 6.34***	0.41 (0.30)	-.45 to 1.00	.21	.45	.65
Conduct composite (<i>n</i> = 191)	1.54 (1.06)	1.07 (0.53)	- 6.22***	0.35 (0.25)	-.31 to 0.89	.19	.38	.51
Anxiety composite (<i>n</i> = 111)	1.38 (1.00)	1.34 (0.66)	- 0.39	0.34 (0.29)	-.66 to .94	.19	.36	.56
Depression composite (<i>n</i> = 83)	1.57 (1.01)	1.28 (0.50)	- 2.86**	0.34 (0.33)	-.52 to .85	.14	.38	.62

^a Each composite was calculated by taking the mean of all relevant items (scored 0 to 6). See Table 1 for the specific items included in each Problem Target composite score

EBP evidence-based practice, *M* mean, *SD* standard deviation

* $p < .05$; ** $p < .01$

of agreement (i.e., characterized as “fair” according to Cicchetti’s (1994) interpretation guidelines). Items with the highest therapist-observer concordance were Time Out ($ICC_{(1,k)} = 0.78$), Trauma Narrative ($ICC_{(1,k)} = 0.76$), and Relaxation ($ICC_{(1,k)} = 0.73$). Items with the lowest concordance were Stimulus/Antecedent Control ($ICC_{(1,k)} = 0.01$), Attending ($ICC_{(1,k)} = -0.01$), Monitoring ($ICC_{(1,k)} = 0.07$), and End of Session Positive ($ICC_{(1,k)} = -0.23$).

Q-correlations were calculated to index item-by-item agreement across the items comprising the Problem Target Composite for a particular session. The mean q-correlation between therapist and observer ratings was 0.37 ($SD = 0.29$), indicating that there was significant overall agreement between therapist and observer ratings as a pattern across all items, even though concordance on some individual items was low (Youngstrom et al. 2000). See Table 3 for q-correlation statistics for each problem target composite.

As seen in Table 2 and Fig. 1a, therapist-report of occurrence of EBP strategies was higher than observers in 2 of the 8 (25%) techniques and 23 of the 24 (96%) content EBP strategies. When a strategy was reported as used (i.e., delivered), the average extensiveness ratings by therapists were higher than observers for most strategies (see Table 2 and Fig. 1b).

Aim 3: Identify Predictors of Observer-Therapist Concordance

To further explore whether certain session, client, or therapist factors may be related to observer-therapist concordance in rating the use of EBP strategies, a linear mixed model was used to predict the q-correlations (an index of observer-therapist agreement). Results for this model are presented in Table 4.

Session and EBP Characteristics

Several session-level variables were found to be significant predictors of observer-therapist concordance. Specifically, the EBP delivered during the session having a structured content ($B = 0.07$, $p < 0.05$) and higher therapist ratings of their ability to carry out intended activities during a session ($B = 0.03$, $p < 0.01$) were both significantly associated with stronger agreement between the observer and therapist.

Client Characteristics

No client characteristics were significantly associated with agreement between the observer and therapist.

Therapist Characteristics

Therapist race/ethnicity was also found to predict observer-therapist concordance. There was significantly lower observer-therapist agreement in EBP strategy use for Hispanic/Latino ($B = -0.11$, $p < 0.01$) or other racial/ethnic minority therapists ($B = -0.16$, $p < 0.01$), compared with Non-Hispanic White therapists.

Therapist attitudes towards EBPs in general were also linked to observer-therapist concordance. Therapists with higher scores regarding divergence, or the perception that EBPs are not as useful as clinical experience tended to have lower levels of agreement between observer and therapist ratings of EBP strategy delivery ($B = -0.05$, $p < 0.01$).

Discussion

This study reported findings from an initial phase of a multi-phased approach to develop a pragmatic measure of EBP strategy delivery within the context of multi-EBP delivery in community children’s mental health services. Specifically,

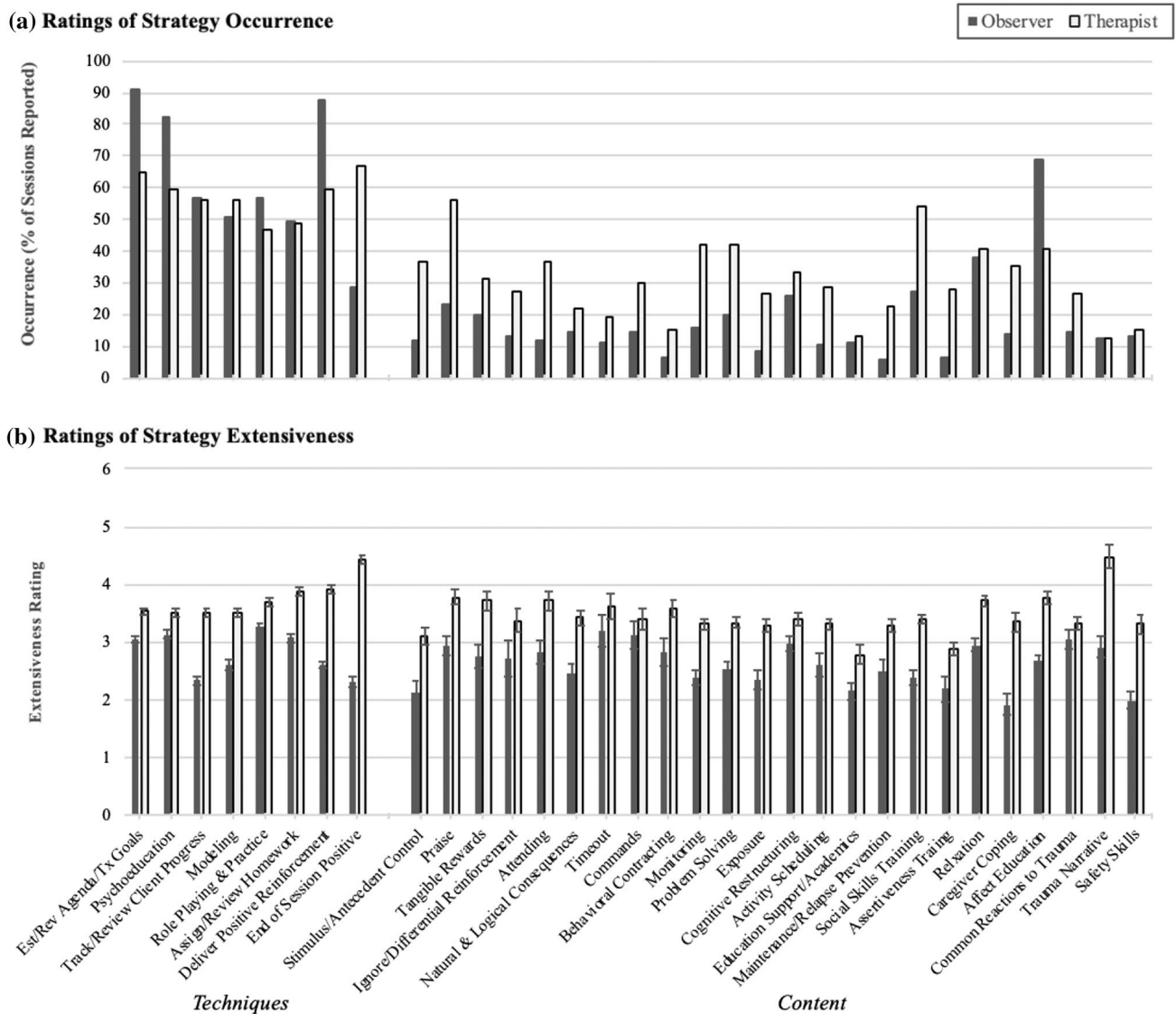


Fig. 1 Observer and therapist ratings of strategy occurrence and extensiveness

patterns of therapist and observer ratings on items assessing EBP delivery were examined, as well as concordance between raters. These findings provide targeted direction for further item refinement and measure development.

How Do Therapists and Observers Differ in Rating Therapist EBP Strategy Delivery?

Descriptive analyses characterizing patterns in individual EBP strategy delivery highlighted differences in breadth (occurrence) versus depth (extensiveness) of EBP strategy delivery based on therapist versus observer report that differed based on EBP strategy type (Content, Techniques). Specifically, for Technique strategies, trained coders endorsed more strategies as having occurred compared to

therapist self-ratings, reflecting *greater breadth*. Observer average extensiveness ratings were lower for individual strategies than compared to therapist self-ratings, reflecting *less depth*. In contrast, for EBP strategies related to a specific content area (e.g., Commands, Cognitive Restructuring), therapists were more likely to rate the occurrence (*greater breadth*) higher compared to observers.

Other studies (Hogue et al. 2014) have found higher concordance between therapists and observers on the targets/content addressed during a session, compared with the treatment techniques. Hogue and colleagues hypothesize that techniques may be more difficult for therapists to reliably report on, as they often co-occur and are multifaceted. In our study, there were several technique items which observers reported as present in a greater percentage of sessions

Table 4 Session, client, and therapist predictors of observer-therapist concordance (Q-correlation)

	Q-correlation	
	<i>B</i>	<i>SE B</i>
Session/EBP variables		
Session EBP structured content	.07*	.04
Session EBP ongoing consultation	-.01	.05
Session conducted in Non-English language	-.06 ⁺	.03
Therapist ability to carry out session intended activities	.03**	.01
Client variables		
Client gender (ref = male)	.03	.02
Client age	-.003	.004
Client race/ethnicity (ref = Non-Hispanic white)		
Hispanic	.05	.06
Other ethnic minority	.04	.06
Therapist variables		
Therapist gender (ref = male)	.07	.05
Therapist race/ethnicity (ref = Non-Hispanic white)		
Hispanic	-.11**	.04
Other ethnic minority	-.16**	.04
Therapist licensed (ref = non licensed)	.01	.04
Therapist years practiced	-.001	.004
Number of EBPs trained in	-.01	.02
Therapist emotional exhaustion	-.001	.02
Therapist general attitudes towards EBP (EBPAS)		
EBPAS divergence	-.05**	.02
EBPAS openness	-.04	.02
Therapist attitudes towards specific EBPs		
Therapist self-efficacy for specific session EBP	.04 ⁺	.02
Perceptions of specific session EBP (PCIS)	-.003	.02

Target of session (i.e., trauma, conduct, anxiety, or depression) was controlled for in all analyses

⁺ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

than did therapists (i.e., Establishing/Reviewing Agenda or Treatment Goals, Psychoeducation, Tracking/Reviewing Client's Progress, Role Play and Practice, Assigning/Reviewing Homework, Delivering Positive Reinforcement and Rewards). This was in contrast to the content items, which the therapists tended to report as being present in sessions more often than observers. Thus, it is possible that therapists may have difficulty recognizing when they are using certain technique strategies.

Which Strategies are More Challenging for Therapist-Observer Agreement?

Results indicate that therapists-observer agreement on therapist delivery of EBP strategies varied by strategy. Concordance between therapist and observer reports was at least

fair for over half of the ECCA- α items. Although it may be tempting to discount therapist report based on this variability in concordance, we believe there is value in pursuing therapist report by examining patterns of lower concordance to inform specific directions for further instrument refinement to produce a pragmatic assessment of EBP delivery in routine care mental health settings across multiple EBPs (see Next Steps below).

Consistent with Garland and colleagues (2010) patterns of inter-rater concordance, all of the items with observer-rated occurrence rates at 10% or less (i.e., infrequently observed items) had low therapist-observer concordance (i.e., ICCs < 0.40) (Assertiveness Training [6%], Maintenance/Relapse Prevention [6%], Behavioral Contracting [6%]; Exposure [8%]; Activity Scheduling [10%]). The remaining low-concordance items were those that may be more challenging for an observer to classify without access to a therapist's intentions (Modeling, Tracking/Reviewing Client Progress, Delivering Positive Reinforcement and Rewards, End of Session Positive, Stimulus/Antecedent Control, Attending, Monitoring, Education support/academics, Caregiver Coping, Communication/Social skills). It is important to note that observers were instructed to refrain from inferring the therapist's intentions when assigning their ratings. Thus, there may be instances where a therapist has full knowledge of their strategy use but the coder has insufficient observable information to code it. Instructions for therapists emphasizing rating their behaviors (vs. intentions or planned activities) may be particularly relevant for these items with lower concordance.

When Do Therapists and Observers Have Stronger Overall Agreement?

Several sessions and therapist characteristics were associated with overall patterns of therapist-observer concordance. Greater cross-method concordance was found for sessions in which an EBP with structured content was delivered, and sessions in which therapists reported they were able to carry out intended activities. Greater therapist-observer agreement was found for therapists who had more positive views towards EBPs in general, and for non-Hispanic White therapists (compared to Latinx and other ethnic minority therapists). These findings build upon previous studies showing relations between therapist training background and therapist reporting accuracy (Hogue et al. 2015; Loades and Myles 2016), and suggest that when therapists felt more mastery over the conduct of the session they evinced stronger reliability with observers. It is unclear what might explain racial/ethnic differences in concordance, but if replicated it would be critical to determine whether this reflects bias in observational rating of non-White therapists' psychotherapy delivery or bona fide differences in self-report reliability. Previous

findings (Lau et al. 2004) suggest that Latinx therapists in the current context report more augmenting adaptations to EBPs to improve their fit for clients compared to White therapists, which may reduce observers' ability to identify strategy delivery or may impact actual extensiveness of strategy delivery. This is a crucial question within an increasingly diverse community mental health workforce.

Implications and Next Steps

The findings from this initial study of the ECCA- α provides targeted direction on the next phase of the ECCA development study aimed to develop pragmatic quality assessment instruments for use when multiple EBPs are delivered in children's mental health services (NIMH R01 MASKED FOR REVIEW). Following review of the data reported here, we initiated development of a substantially revised version of the ECCA item set (ECCA- β). Through focus groups and "think-aloud" interviews, we gathered end user feedback to substantially revise the therapist report measure in an effort to increase concordance between therapist and observer ratings, particularly for lower performing (i.e., low therapist-observer concordance) items. Specifically, therapists provided feedback to adapt the instruction set, rating anchors, item definitions and labels, and user interface to optimize utility and clarity. Major revisions included (1) better aligning the therapist and observer item definitions to help orient therapists to report on observable behaviors, (2) providing examples of high extensiveness ratings for items where therapists typically over-report strategy use related to coders, and (3) adding practice items to help anchor the therapist ratings. The ECCA- β also includes items pertinent to three additional EBPs based on review of intervention materials and expert input from developers and practice experts. Examination of concordance for the revised ECCA- β is currently underway. Based on the findings of the ongoing study, items for which concordance is not improved will be eliminated.

The final measure will also rely on isolating those content and technique strategies that function as quality indicators predictive of client outcomes. It is currently unknown which discrete treatment strategies that are shared across interventions may be responsible for producing the effects generated by EBPs. Following revisions based on the current data and a subsequent concordance test, we will be positioned to conduct a predictive validity test to identify a subset of items that predict client-level outcomes. This work stands to move the common elements approach forward in new ways. Identified quality indicators may subserve efficient quality assurance procedures and advance our understanding of mechanisms of change in children's mental health interventions. Thus, the final product will eventually consist of a smaller number of items that therapists have been shown to reliably

report on, and that predict client outcomes. This iterative measure refinement process aligns with the stakeholder-engaged process and recommended criteria for pragmatic implementation measures set forth in the Psychometric and Pragmatic Evidence Rating Scale (PAPERS; (Stanick et al. 2019).

Strengths and Limitations

One of the primary strengths of this study is to illuminate methods and lessons learned in the development of therapist-report items to measure EBP strategy delivery, particularly within the growing trend of multiple EBP implementation contexts. Balanced with this strength are associated limitations. First, the items included in the ECCA- α as a measure of EBP concordant care are yoked to the specific referent EBPs for specific problem targets. Given the observed challenges with concordance, a useful tool for quality monitoring may only emerge if revised versions yield evidence supporting concordance and predictive validity. Second, even then, the ECCA cannot and is not intended to substitute for protocol-specific measures of treatment fidelity which remain integral for performance feedback in training for specific EBPs. Third, our calculations of observer inter-rater reliability and therapist-observer reliability may underestimate the true reliability, as assignment of coders was not a fully crossed design (Hallgren 2012). Fourth, this study focused on the reliability of therapist report based on associations with observer report and this does not equate to an assessment of the accuracy of therapist-report of EBP strategy delivery. The means of therapist-reported items are not readily interpretable and are reported in this manuscript only for descriptive purposes to compare against observer mean ratings. Finally, the ECCA- α reported in this study does not meet expectations for a pragmatic measure for adoption in that it is lengthy and yet to be validated. However, the objective is to use the findings from the ECCA- α analyses to inform development of a tool aligned with the goals pragmatic measurement.

Conclusions

As multiple EBP implementation initiatives become more common in an attempt to improve children's access to evidence-based community mental health care, the need to assess therapist EBP delivery has become increasingly important. Key requirements of this assessment are that it be pragmatic (i.e., feasible), as well as useful in a multiple EBP context. The current study offers initial steps towards the development of such a measure by examining therapist and observer ratings on items assessing EBP strategy delivery in the context of a multiple EBP implementation effort.

Importantly, the current study does not make claims about reliability and validity of the therapist report measure, but rather aims to understand patterns of therapist-observer concordance and determinants of low reliability with the end goal of identifying a smaller item pool that therapists have been shown to reliably report on, and that predict client care outcomes.

Funding This study was funded by the National Institute of Mental Health (R01MH100134).

Compliance with Ethical Standards

Conflict of Interest Dr. Ann Garland is a co-author on this manuscript and an Associate Editor of Administration and Policy in Mental Health Services Research. Dr. Garland will not be involved in the review of this submission. The authors declare that they have no other conflict of interest.

Ethical Approval All procedures performed in this study that involved human participants were in accordance with the ethical standards of the Institutional Review Boards at the participating academic institutions and practice organizations, and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

- Aarons, G. A. (2004). Mental health provider attitudes toward adoption of evidence-based practice: The Evidence-Based Practice Attitude Scale (EBPAS). *Mental Health Services Research, 6*(2), 61–74. <https://doi.org/10.1023/B:MHSR.0000024351.12294.65>.
- Barnett, M., Brookman-Frazee, L., Regan, J., Saifan, D., Stadnick, N., & Lau, A. S. (2017). How intervention and implementation characteristics relate to community therapists' attitudes toward evidence-based practices: A mixed methods study. *Administration and Policy in Mental Health and Mental Health Services Research, 44*(6), 824–837. <https://doi.org/10.1007/s10488-017-0795-0>.
- Bearsley-Smith, C., Sellick, K., Chesters, J., Francis, K., & Gippsland Adolescent Depression Research Group. (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>.
- Beidas, R. S., Aarons, G., Barg, F., Evans, A., Hadley, T., Hoagwood, K., et al. (2013). Policy to implementation: Evidence-based practice in community mental health – study protocol. *Implementation Science, 8*(1), 38. <https://doi.org/10.1186/1748-5908-8-38>.
- Borntrager, C. F., Chorpita, B. F., Orimoto, T., Love, A., & Mueller, C. W. (2015). Validity of clinician's self-reported practice elements on the Monthly Treatment and Progress Summary. *The Journal of Behavioral Health Services & Research, 42*(3), 367–382. <https://doi.org/10.1007/s11414-013-9363-x>.
- Boustani, M. M., Gellatly, R., Westman, J. G., & Chorpita, B. F. (2017). Advances in cognitive behavioral treatment design: Time for a glossary. *The Behavior Therapist, 40*(6), 199–208.
- Borntrager, C., Chorpita, B., Orimoto, T., Love, A., & Mueller, C. (2013). Validity of clinicians' self-reported practices on the Monthly Treatment and Progress Summary. *Journal of Behavioral Health Services and Research, 1*–15.
- Brookman-Frazee, L., Drahota, A., & Stadnick, N. (2012). Training community mental health therapists to deliver a package of evidence-based practice strategies for school-age children with autism spectrum disorders: A pilot study. *Journal of Autism and Developmental Disorders, 42*(8), 1651–1661. <https://doi.org/10.1007/s10803-011-1406-7>.
- 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*(5), 581–587. <https://doi.org/10.1017/S1352465808004438>.
- Carroll, K., Nich, C., & Rounsaville, B. (1998). Utility of therapist session checklists to monitor delivery of coping skills treatment for cocaine abusers. *Psychotherapy Research, 8*(3), 307–320. <https://doi.org/10.1080/10503309812331332407>.
- Child and Adolescent Mental Health Division. (2003). *Service provider Monthly Treatment and Progress Summary*. Hawaii Department of Health Child and Adolescent Mental Health Division.
- Cho, E., Wood, P. K., Hausman, E. M., Andrews, J. H., & Hawley, K. M. (2019). Evidence-based treatment strategies in youth mental health services: Results from a national survey of providers. *Administration and Policy in Mental Health, 46*, 71–81. <https://doi.org/10.1007/s10488-018-0896-4>.
- Chorpita, B. F., Bernstein, A., & Daleiden, E. L. (2011). Empirically guided coordination of multiple evidence-based treatments: An illustration of relevance mapping in children's mental health services. *Journal of Consulting and Clinical Psychology, 79*(4), 470–480. <https://doi.org/10.1037/a0023982>.
- Chorpita, B. F., Bernstein, A., Daleiden, E. L., & The Research Network on Youth Mental Health. (2008). Driving with roadmaps and dashboards: Using information resources to structure the decision models in service organizations. *Administration and Policy in Mental Health and Mental Health Services Research, 35*(1–2), 114–123. <https://doi.org/10.1007/s10488-007-0151-x>.
- Chorpita, B. F., & Daleiden, E. L. (2009). Mapping evidence-based treatments for children and adolescents: Application of the distillation and matching model to 615 treatments from 322 randomized trials. *Journal of Consulting and Clinical Psychology, 77*(3), 566–579. <https://doi.org/10.1037/a0014565>.
- Chorpita, B. F., & Daleiden, E. L. (2018). Coordinated strategic action: Aspiring to wisdom in mental health service systems. *Clinical Psychology: Science and Practice, 25*(4), e12264. <https://doi.org/10.1111/cpsp.12264>.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. <https://doi.org/10.1037/1040-3590.6.4.284>.
- Cook, J. M., Thompson, R., & Schnurr, P. P. (2015). Perceived Characteristics of Intervention Scale: Development and psychometric properties. *Assessment, 22*(6), 704–714. <https://doi.org/10.1177/1073191114561254>.
- Decker, S. E., & Martino, S. (2013). Unintended effects of training on clinicians' interest, confidence, and commitment in using motivational interviewing. *Drug & Alcohol Dependence, 132*(3), 681–687. <https://doi.org/10.1016/j.drugalcdep.2013.04.022>.
- Dopp, A. R., Hanson, R. F., Saunders, B. E., Dismuke, C. E., & Moreland, A. D. (2017). Community-based implementation of trauma-focused interventions for youth: Economic impact of the learning collaborative model. *Psychological Services, 14*(1), 57–65. <https://doi.org/10.1037/ser0000131>.
- Fung, J. J., & Lau, A. S. (2010). Factors associated with parent-child (dis)agreement on child behavior and parenting problems in Chinese immigrant families. *Journal of Clinical Child & Adolescent Psychology, 39*(3), 314–327. <https://doi.org/10.1080/15374411003691693>.

- Garland, A. F., Bickman, L., & Chorpita, B. F. (2010a). Change what? Identifying quality improvement targets by investigating usual mental health care. *Administration and Policy in Mental Health and Mental Health Services Research*, 37(1–2), 15–26. <https://doi.org/10.1007/s10488-010-0279-y>.
- Garland, A. F., Brookman-Frazee, L., Hurlburt, M. S., Accurso, E. C., Zoffness, R. J., Haine-Schlagel, R., et al. (2010b). Mental health care for children with disruptive behavior problems: A view inside therapists' offices. *Psychiatric Services*, 61(8), 788–795. <https://doi.org/10.1176/ps.2010.61.8.788>.
- Garland, A. F., Brookman-Frazee, L., & McLeod, B. D. (2008a). *Scoring manual for the PRAC study therapy process observational coding system for child psychotherapy: Strategies scale*. Department of Psychiatry: University of California, San Diego.
- Garland, A. F., Hawley, K. M., Brookman-Frazee, L., & Hurlburt, M. S. (2008b). Identifying common elements of evidence-based psychosocial treatments for children's disruptive behavior problems. *Journal of the American Academy of Child & Adolescent Psychiatry*, 47(5), 505–514. <https://doi.org/10.1097/CHI.0b013e31816765c2>.
- Glasgow, R. E., & Riley, W. T. (2013). Pragmatic measures: What they are and why we need them. *American Journal of Preventive Medicine*, 45(2), 237–243. <https://doi.org/10.1016/j.amepr.2013.03.010>.
- Glisson, C., Landsverk, J., Schoenwald, S., Kelleher, K., Hoagwood, K. E., Mayberg, S., Green, P., & The Research Network on Youth Mental Health. (2008). Assessing the Organizational Social Context (OSC) of mental health services: Implications for research and practice. *Administration and Policy in Mental Health and Mental Health Services Research*, 35(1–2), 98–113. <https://doi.org/10.1007/s10488-007-0148-5>.
- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), 23–34. <https://doi.org/10.20982/tqmp.08.1.p023>
- Herschell, A. D., Quetsch, L. B., & Kolko, D. J. (2019). Measuring adherence to key teaching techniques in an evidence-based treatment: A comparison of caregiver, therapist, and behavior observation ratings. *Journal of Emotional and Behavioral Disorders*. <https://doi.org/10.1177/1063426618821901>.
- Hogue, A., Dauber, S., Henderson, C. E., & Liddle, H. A. (2014). Reliability of therapist self-report on treatment targets and focus in family-based intervention. *Administration and Policy in Mental Health and Mental Health Services Research*, 41(5), 697–705. <https://doi.org/10.1007/s10488-013-0520-6>.
- 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 Research*, 42(2), 229–243. <https://doi.org/10.1007/s10488-014-0548-2>.
- 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 Research*, 37(3), 230–244. <https://doi.org/10.1007/s10488-009-0251-x>.
- Kelley, S. D., & Bickman, L. (2009). Beyond outcomes monitoring: Measurement Feedback Systems (MFS) in child and adolescent clinical practice. *Current Opinion in Psychiatry*, 22(4), 363–368. <https://doi.org/10.1097/YCO.0b013e32832c9162>.
- 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 Research*, 37(3), 287–300. <https://doi.org/10.1007/s10488-010-0298-8>.
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>.
- Lau, A. S., & Brookman-Frazee, L. (2016). The 4KEEPS study: Identifying predictors of sustainment of multiple practices fiscally mandated in children's mental health services. *Implementation Science*, 11(1), 31–38. <https://doi.org/10.1186/s13012-016-0388-4>.
- Lau, A. S., Garland, A. F., Yeh, M., McCabe, K. M., Wood, P. A., & Hough, R. L. (2004). Race/ethnicity and inter-informant agreement in assessing adolescent psychopathology. *Journal of Emotional and Behavioral Disorders*, 12(3), 145–156. <https://doi.org/10.1177/10634266040120030201>.
- 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(2), 181–193. <https://doi.org/10.1080/10503300802688460>.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30–46. <https://doi.org/10.1037/1082-989X.1.1.30>.
- McLeod, B. D., Southam-Gerow, M. A., Tully, C. B., Rodriguez, A., & Smith, M. M. (2013). Making a case for treatment integrity as a psychosocial treatment quality indicator for youth mental health care. *Clinical Psychology: Science and Practice*, 20(1), 14–32. <https://doi.org/10.1111/cpsp.12020>.
- 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>.
- Nakamura, B. J., Chorpita, B. F., Hirsch, M., Daleiden, E., Slavin, L., Amundson, M. J., et al. (2011). Large-scale implementation of evidence-based treatments for children 10 years later: Hawaii's evidence-based services initiative in children's mental health. *Clinical Psychology: Science and Practice*, 18(1), 24–35. <https://doi.org/10.1111/j.1468-2850.2010.01231.x>.
- National Institute of Mental Health. (2017). *RFA-MH-17-500: Pragmatic strategies for assessing psychotherapy quality in practice (ROI)*. <https://grants.nih.gov/grants/guide/rfa-files/rfa-mh-17-500.html>
- Orimoto, T. E., Higa-McMillan, C. K., Mueller, C. W., & Daleiden, E. L. (2012). Assessment of therapy practices in community treatment for children and adolescents. *Psychiatric Services*, 63(4), 343–350. <https://doi.org/10.1176/appi.ps.201100129>.
- Regan, J. M. (2014). *Client report of session content in an effectiveness trial: In search of efficient fidelity measurement* [University of California, Los Angeles]. <https://www.escholarship.org/uc/item/2np6k11r>
- 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 Research*, 38(1), 32–43. <https://doi.org/10.1007/s10488-010-0321-0>.
- Stanick, C. F., Halko, H. M., Nolen, E. A., Powell, B. J., Dorsey, C. N., Mettert, K. D., et al. (2019). Pragmatic measures for implementation research: Development of the Psychometric and Pragmatic Evidence Rating Scale (PAPERS). *Translational Behavioral Medicine*. <https://doi.org/10.1093/tbm/ibz164>.
- Ward, A., Regan, J., Chorpita, B., Starace, N., Rodriguez, A., Okamura, K., et al. (2012). Tracking evidence-based practice with youth:

Validity of the MATCH and standard manual consultation records. *Journal of Clinical Child and Adolescent Psychology*. <https://doi.org/10.1080/15374416.2012.700505>.

Youngstrom, E., Loeber, R., & Stouthamer-Loeber, M. (2000). Patterns and correlates of agreement between parent, teacher, and male adolescent ratings of externalizing and internalizing problems. *Journal of Consulting and Clinical Psychology*, 68(6), 1038–1050. <https://doi.org/10.1037//0022-006X.68.6.1038>.

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