The Exposure Guide: A Practical Measure of Exposure Quality

Kristen M. Benito
Jennifer A. Herren
Lesley A. Norris
Kristen M. Gardiner
Molly Choate
Jennifer B. Freeman

Bradley Hospital, East Providence, RI
The Warren Alpert Medical School of Brown University

Quality monitoring is essential for the use of evidence-based interventions (EBIs) in both practice and research settings, yet few quality measures have been developed or validated for any treatment. This study examined the initial psychometric properties of a brief, practical measure of quality for exposure therapy (Exposure Guide; EG) in a sample of youth from three randomized clinical trials for pediatric OCD ($N = 103$ patients and 368 sessions). The EG was initially developed based on the behavioral principles underlying exposure, delivery factors linked to clinical outcomes in prior literature, and with input from both exposure therapy researchers and partners in community mental health settings. Results indicated good to excellent inter-rater reliability (item ICCs = .64 to 1.00). When compared against a validated, time-intensive coding system, each EG item exhibited large correlations with parallel coding system variables; these were significantly larger than correlations with other variables. Variance components analysis demonstrated EG subscale variability at the level of therapists, patients, and time. The EG demonstrates strong initial reliability and construct validity in a clinical trial context; future studies will be needed to establish psychometric properties in practice settings and to elucidate therapist, patient, and treatment course factors that may influence quality.

QUALITY MONITORING and support is essential to use of evidence-based interventions (EBIs) in service settings (Fixsen et al., 2009; Proctor et al., 2011; Schoenwald, 2011). It can facilitate practice sustainment and yield organizational benefits such as reduced staff turnover (Aarons et al., 2009; Stirman et al., 2016), and has been named a key priority in federal legislation like the Children's Health Insurance Program Reauthorization Act (CHIPRA) and the Patient Protection and Affordable Care Act. It is also critically important for rigor and reproducibility in the context of clinical trial research, where it can be used to set training standards, correct delivery "drift," measure detailed components of interventions, and facilitate comparison of the same treatment across different research groups (Baer et al., 2007). The term *quality* refers to the way in which an intervention is delivered to maximize patient improvement. Quality is distinct from, but overlaps with, other constructs that assess treatment delivery, such as adherence, fidelity, and competence. Adherence refers to the degree to which a therapist uses various manual-specified treatment "ingredients" (Perepletchikova et al., 2007). Fidelity captures whether the intervention was delivered as intended by the developer to ensure a valid and reliable test of the intervention under study (Bellg

et al., 2004). Competence is a broader, more subjective construct tapping therapist skill and judgment in manual application (Perepletchikova et al., 2007).

Before quality monitoring can be implemented in research or practice, psychometrically rigorous quality assessments are needed. Despite associated benefits, few quality measures exist for use in either setting, greatly limiting knowledge about quality of treatment delivery (Gearing et al., 2011). While most treatment manuals have accompanying adherence or fidelity checklists (Mowbray et al., 2003), less than 10% of clinicians utilize specific treatment manuals in their practice (Becker et al., 2013) and measures of delivery that do not rely on a specific treatment manual or "ingredient" checklist are rare (Dusenbury et al., 2003). "Gold standard" assessment would involve use of observational coding systems, such as the Therapy Process Observational Coding System (TPOCS; McLeod & Weisz, 2010), with multiple raters trained to reliability. However, such procedures are costly and laborious in both research and practice settings (Perepletchikova et al., 2007).

Quality measures that capture detailed, concrete behaviors have the greatest potential to facilitate EBI translation into practice (Henggeler et al., 1997) by clearly articulating replicable therapist behaviors that can be compared across settings and providing actionable feedback. Those that also measure treatment mechanism—the specific process through which treatment produces improvement—could add further value by providing a proximal therapeutic goal and guiding the use of the most potent prescribed/proscribed behaviors (i.e., what the therapist should and should not do). Theoretically, higher-quality therapy would maximize positive outcomes by better engaging the treatment's hypothesized mechanism of action. A mechanistic emphasis also has potential to facilitate wider implementation by measuring core change elements (e.g., extinction learning) that transcend individual treatment protocols and are applicable across a variety of populations. Despite the promise of this approach, mechanistic measures have largely not been translated into practice settings, and detailed behavioral assessment can be very resource intensive—which is a poor match for busy clinical settings. Thus, we are in need of quality tools that are pragmatic, behaviorally anchored, and psychometrically rigorous.

Exposure therapy is a highly efficacious and broadly applicable EBI (Abramowitz et al., 2019; Higa & Chorpita, 2008) and is an ideal match for mechanism-informed quality measurement. Exposure is the first-line treatment for anxiety and obsessive-compulsive disorder (OCD; Hofmann & Smits, 2008; "Practice Parameter for the Assessment and Treatment of Children and Adolescents with Obsessive-Compulsive Disorder," 2012) and the most common practice element across treatment protocols for anxiety (Chorpita & Daleiden, 2009). The mechanistic theory of exposure is well-articulated and grounded in an understanding of how anxiety develops and is maintained over time. Behavioral theory suggests that maladaptive anxiety is negatively reinforced via the immediate reduction in symptoms that occurs following avoidance (e.g., distraction, accommodation, rituals). Exposure counteracts this cycle by encouraging gradual approach of anxiety-provoking stimuli in the absence of avoidance strategies so that patients can experience new learning (Craske et al., 2014). Changes in distress that occur during exposure are thought to signal and/or facilitate this learning (i.e., fear activation, fear reduction, or fear variability; Benito et al., 2018; Craske et al., 2014). Assessment of fear change is commonly used within exposure tasks as part of clinical treatment and may be practically advantageous for identifying mechanism engagement in training and practice. Despite this, very few studies have examined multiple types of within-exposure fear change over a full course of clinical treatment, and translational studies report inconsistent relationships between various fear changes and clinical outcomes. Aiming to address these gaps and several measurement limitations, the Exposure Process Coding System (EPCS), a microanalytic, time-stamped coding system, captures various aspects of exposure therapy process (including fear changes as well as therapist and patient behaviors) as they occur in time on session recordings (Benito et al., 2012). In a sample of youths who received exposure in one of three randomized controlled trials (Franklin et al., 2011; Freeman et al., 2014; Pediatric OCD Treatment Study [POTS] Team, 2004), findings indicated that greater within-session extinction (WSE; fear reduction observed to occur "on its own") predicted better clinical outcomes, whereas other forms of fear reduction (i.e., those occurring in response to contraindicated antecedents such as avoidance or attempts to "make" fear go down) did not (Benito et al., 2018). This way of measuring short-term fear extinction differs from the traditional approach (often labeled as "within-session habituation") and is consistent with most mechanistic theories of exposure. It is also in line with

recent calls for studies that incorporate an integrated theoretical understanding of exposure mechanism (rather than focus on only one mechanistic theory or on the differences among them; Benito et al., 2024). Other types of fear changes (fear activation and fear variability) did not predict outcomes, suggesting that this WSE metric may be uniquely useful as a "read-out" measure of mechanistic learning and a proximal outcome in quality measurement.

Theory predicts that events during exposure functioning to increase or maintain anxiety (e.g., encouraging approach behavior) will facilitate mechanistic learning and lead to improved treatment outcome, while events functioning to decrease anxiety (e.g., accommodation) during exposure will interfere with learning and be associated with poorer treatment outcome. Measuring these events with EPCS in a sample of young children ($n = 18$) with OCD showed that anxiety increasing behaviors predicted improved outcome at posttreatment and follow-up, while results were in the opposite direction for anxiety decreasing behaviors (Benito et al., 2012). Similarly, in a larger sample of youth with OCD ($N = 111$), EPCS-measured WSE significantly predicted outcome on all measures, and therapist behaviors predicted treatment response as well as WSE—suggesting that even among closely monitored and highly adherent clinical trial therapists, measuring quality can add predictive value (Benito et al., 2018; Benito, Machan, et al., 2021). Moreover, evidence from a community-based randomized training trial shows that therapists who were trained to optimize EPCS-rated quality indicators met or exceeded quality benchmarks based on clinical trial therapists, and they had higher rates of patient WSE and treatment response (vs. those who received standard training; Benito, Herren, et al., 2021). Together, these findings support a causal relationship between EPCS-rated quality indicators (in the form of specific therapist behaviors) and clinical outcomes of exposure therapy in both research and clinical practice settings.

Despite promising findings for EPCS-rated quality indicators in exposure therapy, it is time-intensive, requires extensive training and specialized software, and is infeasible for use in practice and most research settings. The need for pragmatic quality measurement strategies is particularly relevant for exposure given that it is associated with unique barriers (e.g., therapist negative beliefs; Deacon et al., 2013a, 2013b) and rarely used in practice settings (Becker et al., 2004; Whiteside et al., 2016; Wolitzky-Taylor et al., 2015). When exposure is used, delivery typically deviates from specialist recommendations. For example, therapists report using anxiety decreasing behaviors frequently (e.g., distraction, 81.3%; thought-stopping, 84.7%; Whiteside et al., 2016)—in contrast to the anxiety-increasing or maintaining behaviors advocated by anxiety specialists (Stewart et al., 2016) and supported by mechanistic theory. A low-burden measure of exposure quality would be a valuable tool for supporting exposure implementation and quality improvement in practice settings, and for facilitating rigor and reproducibility in clinical trials. However, studies have yet to develop psychometrically sound measures of exposure quality for use in research or practice settings.

The goal of the present study is to examine the psychometric properties of a brief, practical measure of exposure quality (Exposure Guide; EG) in a sample of youth from three randomized clinical trials for pediatric OCD. The EG was designed to capture EPCS quality indicators, and to be used as a low-burden quality monitoring and training tool that is relevant for treatment outcome and matches the needs of practice settings. Specifically, this study aimed to establish initial reliability and construct validity of the EG against the original, time-intensive EPCS coding system.

Methods

PARTICIPANTS

Participants ($N = 103$) were 5-18 years of age ($M = 10.12$, $SD = 3.59$), 54.1% female, 91.2% White, 3.5% Black, 4.4% other/more than one race, and 5.3% Latine. All participants had a primary diagnosis of OCD and received exposure-based cognitive behavioral therapy (CBT) in one of three randomized controlled trials (Pediatric OCD Treatment Study "POTS" trials). To be included in the current study, youth had (a) been randomized to a study condition with CBT and (b) at least two recorded exposure sessions. Eligible youth from POTS I (POTS Team, 2004) were 7–17 years old and received CBT alone or with a serotonin reuptake inhibitor (SRI). Eligible youth from POTS II (Franklin et al., 2011) were 7–17 years old, considered to be partial SRI responders, and received CBT with an SRI. Eligible youth from POTS Jr (Freeman et al., 2014) were 5–8 years old and received CBT alone.

EXPOSURE THERAPY

Exposures were administered by trained study therapists ($N = 26$) who attended weekly supervision and followed published CBT treatment manuals (POTS I and II: March & Mulle, 1998;

POTS Jr.: Freeman & Garcia, 2008) to which they were highly adherent (Franklin et al., 2011; Freeman et al., 2014; POTS Team, 2004). These manuals detail treatment ingredients including exposure (11 sessions in POTS I and II and 8 sessions in POTS Jr.), but delivery approach was not explicitly described in the manuals (e.g., exposure intensity, therapist behaviors during exposure, SUDS sampling).

SESSION DATA

Study measures were rated using session videos collected during the original trials. Session videos were sampled to equally represent early, middle, and late protocol exposures for each participant. Sessions were not rated if the recording was missing or damaged or the exposure occurred off camera (e.g., out of the office). We rated 42% of all exposure sessions ($N = 368$; 32.1% early, 33.4% middle, and 26.1% late protocol sessions).

EG DEVELOPMENT AND DESCRIPTION

The EG was initially designed for CMH therapists as a pragmatic quality feedback tool to enhance transdiagnostic exposure therapy training. Consistent with the principles of pragmatic measurement (Glasgow & Riley, 2013), EG development placed particular emphasis on being valuable for community partners, low burden, clinically actionable, and relevant for treatment outcome. Initial EG items were drafted based on parallel codes from EPCS that had demonstrated predictive value with treatment outcome. We made multiple rounds of revisions using feedback about optimal utility from community partners in two CMH agencies, feedback about theoretical fidelity from three exposure therapy researchers, and iterative pilot testing by EG developers. This co-design approach is one in which investigators contribute treatment knowledge while community partners contribute local expertise to adapt process and content based on context variables that investigators cannot anticipate (Grindell et al., 2022; Vargas et al., 2022). The resulting revisions focused on both EG content (e.g., to ensure user-friendly wording, use of response scales with a limited range, retention of items with high perceived utility) and format (e.g., visually grouping items to reduce burden and enhance understanding). The final EG includes items that assess therapist behaviors, patient learning, and overall exposure difficulty (see Table 1 for individual item descriptions and response scales). Therapist behavior items are grouped into functionally defined subscales (increasing, decreasing, or neutral short-term effect on fear); items were assigned to a subscale based on strong theoretical rationale (e.g., relaxation in the fear decreasing category) and consensus among exposure specialists. Importantly, EG development emphasized relevance for any disorder commonly treated with exposure (e.g., social anxiety disorder, generalized anxiety disorder, illness anxiety disorder). We note that the EG includes several additional items that were designed for utility in a practice setting (e.g., assessment of barriers when exposure did not take place); the present study did not include these items because (a) they have limited relevance in the context of high protocol adherence in a clinical trial, and (b) a parallel item does not exist in EPCS for comparison. The EG is available from the corresponding author upon request.

CONSTRUCT VALIDITY MEASURE: EXPOSURE PROCESS CODING SYSTEM (EPCS)

EPCS is a microanalytic coding system that uses Noldus Observer software to link coded observations with video and generates a timestamp for each variable. EPCS continuously measures therapist, parent (if present), and patient behaviors during exposure. Therapist behavior variables (duration of each, described in Table 1) were of primary interest in this study. EPCS also measures patient-reported subjective units of distress (SUDS), which was used to calculate whether therapists asked for SUDS ratings (yes/no). EPCS further assesses observer-rated patient fear level (continuously measured on a 0 to 5 scale); these were used to calculate whether high peak fear (score of 4–5) or low peak fear (score of 0–1) occurred in each exposure (yes/no). To measure patient WSE, coders judged whether each instance of observed fear decrease was associated with an antecedent likely to explain the reduction (e.g., rituals, avoidance, reduction of task difficulty, unplanned change in exposure stimulus, focus on anticipation of exposure end). Decreases not better explained by such antecedents (i.e., occurring "on their own") are more likely to signal an internal learning process and were judged as WSE. Given the complexity of WSE measurement considerations, we refer interested readers to Benito et al. (2018) for in-depth rationale and discussion. EPCS has demonstrated good inter-coder reliability and predictive validity with treatment outcome in the current clinical trial sample (previously published in Benito et al., 2018; Benito et al., 2021a) and in a small sample of young children with OCD (N = 18; Benito, et al., 2012). It has also demonstrated good inter-coder reliability and sensitivity to therapist training effects in two studies of therapists treating youth with anxiety or OCD in

Table 1
Descriptive Statistics and Interrater Reliability for EG Items and Subscales

EG Item/Subscale	Description	Response Scale	M(SD)	ICC
Took Ratings	Therapist took ratings of exposure difficulty (e.g., SUDS ratings)	0 = no	.85(.36)	.94
Too Hard	Overall exposure difficulty was too hard	1 = yes	.11(.31)	1.00
Too Easy	Overall exposure difficulty was too easy		.15(.36)	.64
WSE	Client experienced some within-session extinction (decrease in distress occurring "on its own")		.71(.46)	.75
Encourage Approach	Therapist encouraged the client to approach or focus on the exposure without avoidance	1 = none 2 = some	2.52(.51)	.80
Intensify	Therapist introduced new information or changed the exposure to increase the client's anxiety	3 = a lot	1.54(.58)	.81
Par. Accom.	Therapist discouraged family members from reducing the client's anxiety		1.08(.29)	.83
Teach	Therapist provided education about principles that are consistent with this exposure model		1.55(.59)	.91
Externalize	Therapist referred to anxiety as being separate from the client		1.76(.55)	.96
Change Thoughts	Therapist encouraged the client to change or replace an anxious thought to reduce anxiety		1.35(.61)	.83
Unrelated Talk	Therapist talked about things unrelated to the exposure		1.44(.66)	.92
Relaxation	Therapist encouraged client to use relaxation or imagery		1.03(.17)	1.00
Accomm.	Therapist introduced new information or changed the exposure to reduce the client's anxiety		1.39(.52)	.89
Increasing	Sum of response scores for encourage approach, intensify, and address parent accommodation	Range = 3-9	5.16(1.00)	.86
Neutral	Sum of response scores for teach, externalize, and change anxious thoughts		4.67(1.19)	.93
Decreasing	Sum of response scores for Unrelated talk, relaxation, and accommodation		3.86(.95)	.89

CMH settings (Benito et al., 2021a, 2021b, 2021c).

EPCS CODER AND EG RATER TRAINING
EPCS coders were four bachelor's-level research assistants and one postdoctoral fellow. Initial training included guided reading of the EPCS manual, observing trained coders, coding with supervision of trained coders, and independent coding of training videos to criterion (K or ICC > .70 on all variables). Trained coders attended weekly meetings to discuss EPCS implementation and prevent drift. Ten percent of tapes were double-coded for inter-coder reliability, which was excellent in the present sample (ICC or K range .71–.98). To prevent drift, 10% of tapes were subject to ongoing review by the first author.

EG raters were two Ph.D.-level clinicians who had prior experience providing exposure therapy but were not familiar with the EPCS coding system. Initial EG training included guided manual reading followed by independent coding of training videos to criterion (K or ICC > .70 on all variables). Trained raters attended monthly meetings to discuss EG implementation and prevent drift.

Ten percent of tapes were double-coded for calculating inter-coder reliability (results are presented below). To prevent drift, 10% of tapes were subject to ongoing review by the first author.

Analytic Plan

RELIABILITY
We calculated interrater reliability for each EG item using the ICC (2,2) based on a two-way random effects model (Koo & Li, 2016). This approach estimates reliability using the mean scores across coders and allows for generalizability to other samples. ICCs of .75 and above are thought to reflect "excellent" agreement and those between .60 and .74 reflect "good" agreement (Cicchetti, 1994). Internal consistency of EG subscales (fear-increasing, -neutral, and -decreasing therapist behaviors; see Table 1) was evaluated using Cronbach's alpha.

Construct Validity
To evaluate convergent and divergent validity for each EG item and subscale, we compared the size of its correlation with the parallel EPCS variable

versus with other EPCS variables. Correlation analyses used Spearman's rank order due to the nonnormal response distribution associated with response scales used on the EG. Significance of correlation size differences was evaluated by calculating a z-score (difference score divided by the square root of the pooled variance; Hedges and Olkin, 1985).

Variance Components Analysis
We conducted variance components analysis of the EG subscales (fear-increasing, -neutral, and -decreasing therapist behaviors) to evaluate sources of variability in these scores. Information about variability at each data level (e.g., associated with therapists or patients) can be used to guide future work aiming to improve quality (e.g., by identifying and intervening on therapist- or patient-level factors that influence variability in quality). We calculated variance components for each subscale using a mixed model with restricted maximum likelihood estimation in SPSS for each of the following data levels: study (POTS I, POTS II, or POTS Jr.), therapist, patient (nested within therapist), and time (nested within therapist and patient). Time represents change in scores across treatment (measured by session number). To aid interpretability, variance estimates are presented as a proportion of the total estimated variance.

Results

RELIABILITY

Interrater reliability was excellent for all EG items (ICC range .75 to 1.00; see Table 1, with the exception of the item assessing whether the exposure was "too easy" (ICC = .64 indicating good reliability). Interrater reliability was excellent for EG subscales (ICC range .86 to .93). Internal consistency was low for each of the EG subscales (fear-increasing α = .33, fear-neutral α = .41, fear-decreasing α = .28), indicating that the different therapist behaviors within each functional category are not consistently used together.

CONSTRUCT VALIDITY

Each EG item and subscale exhibited a large correlation with the parallel EPCS variable (Rosenthal & Rosnow, 1984) that was also significantly larger than its correlations with other EPCS variables ($p < .05$ for all z-scores). All correlations are presented in Table 2.

VARIANCE COMPONENTS ANALYSIS

For therapist fear-increasing and -decreasing behavior subscales on the EG, therapists and patients account for similar and substantial proportions of variance (all results presented in Table 3), indicating that use of these behaviors varied across individual therapists as well as individual patients. Time in treatment also accounted for some proportion of variance in fear-decreasing behaviors, suggesting that this changes across treatment. For the fear-neutral therapist behavior subscale, a substantial proportion of variance was accounted for by patients and time, suggesting that therapist use of these behaviors differs by patient and changes across treatment. Study membership did not account for a substan-

Table 2
EG Item and Subscale Correlations With EPCS

		EPCS															
		Took SUDS	High Peak Fear	Low Peak Fear	WSE	Encourage Approach	Intensify	Par. Accom.	Teach	Externalize	Change Thoughts	Unrelated	Relaxation	Accomm.	Increasing	Neutral	Decreasing
Exposure Guide	Took Ratings	.76*	-.04	-.09	.20**	.27**	.13*	.00	.13*	.06	.13*	-.01	.06	-.02	.25**	.15**	-.04
	Too Hard	-.09	.38**	-.10	-.14**	.08	.00	-.02	.06	.07	-.05	.10	.14**	.18**	.07	.04	.16**
	Too Easy	-.08	-.14**	.44**	-.16**	-.29**	-.07	.00	-.16**	-.12*	-.12*	-.09	-.06	-.09	-.27**	-.21**	-.14**
	WSE	.26**	-.04	-.21**	.37**	.21**	.06	.06	.08	.11*	.12*	-.12*	-.08	-.05	.20**	.16**	-.09
	Encourage Approach	.11*	.07	-.13*	.26**	.46**	.34**	.00	-.09	-.05	-.13*	-.26**	-.03	-.02	.48**	-.10	-.21**
	Intensify	.03	.11*	-.10	.20**	.33**	.56**	-.06	.01	.09	-.13*	-.05	.00	.06	.40**	-.01	.01
	Par. Accom.	-.08	.20**	.01	-.13*	.08	.02	.37**	.11*	.04	-.01	.02	-.04	.13*	.09	.07	.09
	Teach	.11*	.14**	-.18**	.11*	.18**	.03	.10	.66**	.45**	.08	.16**	-.03	.17**	.16**	.57**	.20**
	Externalize	.05	.05	-.04	.03	.13*	.03	.01	.33**	.62**	.07	.09	-.01	.16**	.12*	.51**	.13*
	Change Thoughts	.16**	-.05	-.07	.00	.05	-.09	-.03	.13*	.16**	.64**	.02	.05	.21**	.03	.39**	.07
	Unrelated	-.10	.07	-.04	-.18**	-.02	-.10	-.01	.21**	.22**	-.06	.72**	.05	.23**	-.05	.17**	.66**
	Relaxation	.09	.02	-.02	.03	.05	.02	-.03	-.01	-.04	.01	.06	.79**	.00	.05	-.03	.05
	Accomm.	-.06	.29**	-.15**	-.10	.05	-.14	.00	.19**	.24**	.15**	.21**	.07	.55**	.02	.25**	.45**
	Increasing	.06	.16**	-.13*	.24**	.48**	.55**	.07	-.02	.03	-.16**	-.17**	-.03	.05	.54**	-.05	-.09
	Neutral	.16**	.06	-.14**	.07	.17**	-.02	.03	.53**	.57**	.41**	.10	.00	.25**	.15**	.72**	.18**
	Decreasing	-.09	.20**	-.13*	-.16**	.01	-.15**	.01	.26**	.27**	.05	.59*	.17**	.49**	-.03	.25**	.71**

Note. Blue cells indicate correlations between subscales of one measure and the individual items that comprise the subscale on the other measure.*$p < .05$.
**$p < .01$.

Table 3
Variance Components for EG Subscales

EG Subscale	Study	Therapist	Patient	Time
Increasing	.03	.13	.12	.01
Neutral	.03	.04	.31	.31
Decreasing	.00	.10	.11	.08

tial proportion of variance in any subscale, suggesting that therapist behaviors assessed with the EG did not vary across POTS I, POTS II, and POTS Jr.

Discussion

This study examined the psychometric properties of the EG, a brief measure of quality in exposure therapy, within a sample of youth from three prior clinical trials of youth with OCD. Overall, findings highlight the utility of the EG as a psychometrically sound and pragmatic measure of exposure quality. Good to excellent interrater reliability was demonstrated for all EG items. Convergent and divergent validity with EPCS, a validated but time-intensive microanalytic quality measure, was also supported. Demonstration of these properties is an important step in developing a psychometrically sound assessment of exposure quality, particularly as the majority of quality measurements of mental healthcare services to date have not been evaluated along these domains in any context (12% examined for reliability, 3% for validity; Hermann et al., 2000). In addition, the constructs assessed (e.g., encouraging approach of feared stimuli) are guided by the presumed mechanism underlying exposure therapy and have been linked with differential clinical outcomes (Benito et al., 2012; Benito et al., 2018; Benito et al., 2020). Time demands of microanalytic observational coding systems like EPCS, which include time training raters to fidelity, coding videotapes, and interrater reliability assessment, represent a key barrier to their use in regular quality monitoring in both research and practice settings. A tool like the EG may be better suited to meet the needs of both settings.

Despite strong evidence of interrater reliability and construct validity, internal consistency was low within functionally defined EG therapist behavior subscales (i.e., fear-increasing, -neutral and -decreasing behaviors). In line with the principles of pragmatic measurement (Glasgow & Riley, 2013), brevity was prioritized in EG development to avoid redundancy and enhance clinical utility. Given that subscales with a low number of items are more likely to have low internal consistency (Taber, 2018), high alphas were not necessarily

expected but were examined to assess how therapist behaviors within the same domain are used together and inform researcher decisions regarding use of subscale or individual items in subsequent studies. Low internal consistency in this context suggests that functionally similar therapist behaviors were not consistently used together. For example, a therapist may have regularly encouraged clients to approach or focus on exposures, without concurrently working to reduce parental accommodation. This does not change the function of either fear-increasing behavior, but indicates that therapists may deploy only one or two of these strategies rather than using all within a session. As such, low internal consistency likely reflects true variability in the use of techniques within a subscale rather than signaling measurement error. Furthermore, although low reliability can generally interfere with measure validity, internal consistency has limited relevance for understanding validity (McCrae et al., 2011). For example, high internal consistency can reflect item redundancy and/or undue narrowness, thus reducing predictive power (McCrae et al., 2011). Importantly, EG subscales did demonstrate very strong interrater reliability. Researchers using the EG in the future should consider that therapist behaviors within a subscale are not necessarily used together, and this should be weighed in the context of study goals when selecting EG variables of interest (e.g., subscales versus individual items). Finally, subscale internal consistency was likely influenced by some behaviors that were used infrequently (e.g., relaxation). These may be used more often in community settings, and additional work will be needed to examine subscale performance in that context.

Therapist use of fear-increasing and -decreasing behaviors varied substantially by therapist and by patient. Patients were also an important source of variability for fear-neutral therapist behaviors (e.g., teaching, practicing cognitive skills). This is unsurprising, as any therapeutic process is inherently dynamic and will reflect characteristics of both therapists and patients. For example, patients with increased symptom severity might elicit more fear-decreasing behaviors from therapists, while therapists with fewer negative beliefs surrounding exposure use might use more fear-increasing behaviors (Deacon et al., 2013a, 2013b). Therapist fear-decreasing and -neutral behaviors also changed across time in treatment. It is possible that these behaviors were used more often early in treatment to set the stage for future exposure work or that these behaviors moved dynamically with patient symptom change throughout treatment. Notably, we did not observe much variabil-

ity at the trial level, suggesting that the therapist behaviors captured by the EG were not influenced by trial differences in the treatment protocol or other procedures. Future studies should seek to understand specific factors that might contribute to therapist, patient, and time-related variability in therapist delivery of exposure. Better understanding of these factors will help to personalize exposure-based treatments and therapist training moving forward.

There are several limitations of the present study to consider. Perhaps most important, while the EG was developed for utility in both research practice settings, this study took place in a research setting. Findings could differ in practice settings where patient, therapist, and organizational characteristics are more variable and resources are often more limited (e.g., for rater training). As one example, we found limited use of some therapist behaviors (e.g., relaxation), yet relaxation is very commonly used in CMH settings to address anxiety (Whiteside et al., 2016). Similarly, findings may differ when the EG is rated by a therapist rather than a study rater. On one hand, therapist ratings would necessarily be based on recall (rather than tape-watching), possibly reducing accuracy. On the other hand, therapists may have additional case knowledge that facilitates the functional assessment required for EG ratings, possibly improving accuracy. Finally, interrater reliability was somewhat lower—though still strong—for items involving more cognitive content (e.g., changing anxious thoughts) or functional complexity (e.g., WSE); it is possible that more support (e.g., through training) will be important for attaining reliability on these items among therapists that do not specialize in exposure therapy. Future studies will be needed to examine whether the strong psychometric properties of the EG demonstrated in the current study are consistent when rated by therapists and/or in community treatment settings. Critically, our sample included a very small proportion of youth from minoritized racial/ethnic backgrounds. Exposure mechanism and optimal delivery may differ for historically marginalized populations (Benito et al., 2024). For example, evidence suggests that exposure to discrimination increases anxious arousal among Latine individuals (Zvolensky et al., 2022), and that racial differences in threat response are apparent in neural, behavioral, and psychophysiological domains (a finding partially explained by exposure to negative life experiences;

Harnett et al., 2019). Finally, EG scores could not be compared to treatment adherence estimates in this study, as adherence was assessed differently across the three included trials and was generally not sufficiently detailed for comparison (e.g., yes/no item assessing exposure use). This is typical of adherence measures in most trials and further highlights the comparative strengths of process-driven, theoretically grounded quality assessment tools.

Limitations notwithstanding, the current study supports the initial psychometric properties of the EG as a pragmatic tool to better understand exposure quality in the context of clinical trial research. It also has potential for utility in practice settings, though additional research will be needed to determine which end-users can feasibly and reliably use it (e.g., therapists, patients, supervisors), how it fits within different settings (e.g., based on implementation climate), and confirm its psychometric properties. If successful, measures like the EG could be used in multiple ways. For example, therapists might use it as a self-feedback tool and/or to populate key information in session notes, supervisors might use it to streamline case discussion in supervision meetings, or third-party payers might incentivize its use as part of value-based care. Quality measures like the EG may also have important policy implications. Federal legislation such as the Children's Health Insurance Program Reauthorization Act (CHIPRA) and the Patient Protection and Affordable Care Act have prioritized ongoing monitoring and improvement of service quality, leading to various national initiatives designed to provide recommendations for quality measurement. However, only nine of the recommended quality measures resulting from these initiatives were focused on provision of mental health services and none focused on quality of exposure therapy delivery (Zima et al., 2013). The EG could help to fill this gap, while placing comparatively limited time demands on patients and providers to ensure more widespread uptake.

References

Aarons, G. A., Sommerfeld, D. H., Hecht, D. B., Silovsky, J. F., & Chaffin, M. J. (2009). The impact of evidence-based practice implementation and fidelity monitoring on staff turnover: evidence for a protective effect. *Journal of Consulting and Clinical Psychology, 77*(2), 270–280. https://doi.org/10.1037/a0013223.

Abramowitz, J. S., Deacon, B. J., & Whiteside, S. P. H. (2019). *Exposure therapy for anxiety: principles and practice* (2nd ed.). Guilford Publications.

Baer, J. S., Ball, S. A., Campbell, B. K., Miele, G. M., Schoener, E. P., & Tracy, K. (2007). Training and fidelity monitoring of behavioral interventions in multi-site addictions research. *Drug and Alcohol Dependence, 87*(2–3), 107–118. https://doi.org/10.1016/j.drugalcdep.2006.08.028.

Becker, C. B., Zayfert, C., & Anderson, E. (2004). A survey of psychologists' attitudes towards and utilization of exposure therapy for PTSD. *Behaviour Research and Therapy, 42*(3), 277–292. https://doi.org/10.1016/S0005-7967(03)00138-4.

Becker, E. M., Smith, A. M., & Jensen-Doss, A. (2013). Who's using treatment manuals? A national survey of practicing therapists. *Behaviour Research and Therapy, 51*(10), 706–710. https://doi.org/10.1016/j.brat.2013.07.008.

Bellg, A., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., Ogedegbe, G., Orwig, D., Ernst, D., Czajkowski, S., & Treatment Fidelity Workgroup of the NIH Behavior Change Consortium (2004). Enhancing treatment fidelity in health behavior change studies: best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychology, 23*, 443–451. https://doi.org/10.1037/0278-6133.23.5.443.

Benito, K., Pittig, A., Abramowitz, J., Arch, J., Chavira, D., de Kleine, R., De Nadai, A., Hermans, D., Hofmann, S. G., Hoyer, J., Huppert, J. D., Kircanski, K., McEvoy, P. M., Meyer, H., Monfils, M. H., Papini, S., Rief, W., Rosenfield, D., Storch, E. A., Telch, M. W., ... the Exposure Therapy Consortium (2024). Mechanisms of change in exposure therapy for anxiety and related disorders: A research agenda. *Clinical Psychological Science*. https://doi.org/10.1177/21677026241240727.

Benito, K. G., Conelea, C., Garcia, A. M., & Freeman, J. B. (2012). CBT specific process in exposure-based treatments: Initial examination in a pediatric OCD sample. *Journal of Obsessive-Compulsive and Related Disorders, 1*(2), 77–84. https://doi.org/10.1016/j.jocrd.2012.01.001.

Benito, K. G., Herren, J., Freeman, J. B., Garcia, A. M., Kemp, J., Chorpita, B. F., Sung, J., Arora, A., Georgiadis, C., Stewart, E., Wellen, B., Frank, H. E., Schiavone, E., Gervasio, M., & Machan, J. (2021). Improving delivery behaviors during exposure therapy: A cluster randomized controlled training trial with community therapists. *Evidence-Based Practice in Child and Adolescent Mental Health, 6*(2), 173–190. https://doi.org/10.1080/23794925.2021.1917018.

Benito, K. G., Machan, J., Freeman, J. B., Garcia, A. M., Walther, M., Frank, H., Wellen, B., Stewart, E., Edmunds, J., Kemp, J., Sapyta, J., & Franklin, M. (2018). Measuring fear change within exposures: Functionally-defined habituation predicts outcome in three randomized controlled trials for pediatric OCD. *Journal of Consulting and Clinical Psychology, 86*(7), 615–630. https://doi.org/10.1037/ccp0000315.

Benito, K. G., Machan, J., Freeman, J. B., Garcia, A. M., Walther, M., Frank, H., Wellen, B., Stewart, E., Edmunds, J., Sapyta, J., & Franklin, M. E. (2021a). Therapist behavior during exposure tasks predicts habituation and clinical outcome in three randomized controlled trials for pediatric OCD. *Behavior Therapy, 52*(3), 523–538. https://doi.org/10.1016/j.beth.2020.07.004.

Benito, K. G., Herren, J., Freeman, J. B., Garcia, A. M., Block, P., Cantor, E., Chorpita, B. F., Wellen, B., Stewart, E., Georgiadis, C., Frank, H., & Machan, J. (2021b). Improving delivery behaviors during exposure for Pediatric OCD: a multiple baseline training trial with community therapists. *Behavior Therapy, 52*(4), 806–820. https://doi.org/10.1016/j.beth.2020.10.003.

Benito, K. G., Herren, J., Freeman, J. B., Garcia, A. M., Kemp, J., Chorpita, B. F., Sung, J., Arora, A., Georgiadis, C., Stewart, E., Wellen, B., Frank, H. E., Shiavone, E., Gervasio, M., & Machan, J. (2021c). Improving delivery behaviors during exposure therapy: a cluster randomized controlled training trial with community therapists. *Evidence-Based Practice in Child and Adolescent Mental Health, 6*(2), 173–190. https://doi.org/10.1080/23794925.2021.1917018.

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.

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.

Craske, M. G., Treanor, M., Conway, C. C., Zbozinek, T., & Vervliet, B. (2014). Maximizing exposure therapy: An inhibitory learning approach. *Behaviour Research and Therapy, 58*, 10–23. https://doi.org/10.1016/j.brat.2014.04.006.

Deacon, B. J., Farrell, N. R., Kemp, J. J., Dixon, L. J., Sy, J. T., Zhang, A. R., & McGrath, P. B. (2013a). Assessing therapist reservations about exposure therapy for anxiety disorders: the therapist beliefs about exposure scale. *Journal of Anxiety Disorders, 27*(8), 772–780. https://doi.org/10.1016/j.janxdis.2013.04.006.

Deacon, B. J., Lickel, J. J., Farrell, N. R., Kemp, J. J., & Hipol, L. J. (2013b). Therapist perceptions and delivery of interoceptive exposure for panic disorder. *Journal of Anxiety Disorders, 27*(2), 259–264. https://doi.org/10.1016/j.janxdis.2013.02.004.

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. *Health Education Research, 18*(2), 237–256. https://doi.org/10.1093/her/18.2.237.

Fixsen, D. L., Blase, K. A., Naoom, S. F., & Wallace, F. (2009). Core implementation components. *Research on Social Work Practice, 19*(5), 531–540. https://doi.org/10.1177/1049731509335549.

Franklin, M. E., Sapyta, J., Freeman, J. B., Khanna, M., Compton, S., Almirall, D., Moore, P., Choate-Summers, M., Garcia, A., Edson, A. L., Foa, E. B., & March, J. S. (2011). Cognitive behavior therapy augmentation of pharmacotherapy in pediatric obsessive-compulsive disorder: the Pediatric OCD Treatment Study II (POTS II) randomized controlled trial. *JAMA, 306*(11), 1224–1232. https://doi.org/10.1001/jama.2011.1344.

Freeman, J. B., & Garcia, A. M. (2008). *Family based treatment for young children with OCD: therapist guide.* Oxford University Press. https://doi.org/10.1093/med:psych/9780195373639.001.0001.

Freeman, J., Sapyta, J., Garcia, A., Compton, S., Khanna, M., Flessner, C., FitzGerald, D., Mauro, C., Dingfelder, R., Benito, K., Harrison, J., Curry, J., Foa, E., March, J., Moore, P., & Franklin, M. (2014). Family-based treatment of early childhood obsessive-compulsive disorder: the Pediatric Obsessive-Compulsive Disorder Treatment Study for Young Children (POTS Jr)— A randomized clinical trial. *JAMA Psychiatry, 71*(6), 689–698. https://doi.org/10.1001/jamapsychiatry.2014.170.

THE EXPOSURE GUIDE

Gearing, R. E., El-Bassel, N., Ghesquiere, A., Baldwin, S., Gillies, J., & Ngeow, E. (2011). Major ingredients of fidelity: a review and scientific guide to improving quality of intervention research implementation. *Clinical Psychology Review, 31*(1), 79–88. https://doi.org/10.1016/j.cpr.2010.09.007.

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.amepre.2013.03.010.

Grindell, C., Coates, E., Croot, L., & O'Cathain, A. (2022). The use of co-production, co-design and co-creation to mobilise knowledge in the management of health conditions: a systematic review. *BMC Health Services Research, 22*(1), 877. https://doi.org/10.1186/s12913-022-08079-y.

Harnett, N. G., Wheelock, M. D., Wood, K. H., Goodman, A. M., Mrug, S., Elliott, M. N., Schuster, M. A., Tortolero, S., & Knight, D. C. (2019). Negative life experiences contribute to racial differences in the neural response to threat. *NeuroImage, 202*, 116086. https://doi.org/10.1016/j.neuroimage.2019.116086.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego, CA: Academic Press.

Henggeler, S. W., Melton, G. B., Brondino, M. J., Scherer, D. G., & Hanley, J. H. (1997). Multisystemic therapy with violent and chronic juvenile offenders and their families: the role of treatment fidelity in successful dissemination. *Journal of Consulting and Clinical Psychology, 65*(5), 821–833. https://doi.org/10.1037//0022-006x.65.5.821.

Hermann, R. C., Leff, H. S., Palmer, R. H., Yang, D., Teller, T., Provost, S., Jakubiak, C., & Chan, J. (2000). Quality measures for mental health care: results from a national inventory. *Medical Care Research and Review, 57*(Suppl 2), 136–154. https://doi.org/10.1177/1077558700057002S08.

Higa, C. K., & Chorpita, B. F. (2008). Evidence-based therapies: Translating research into practice. In R. G. Steele, T. D. Elkin, & M. C. Roberts (Eds.), *Handbook of evidence-based therapies for children and adolescents: bridging science and practice* (pp. 45–61). US: Springer. https://doi.org/10.1007/978-0-387-73691-4_4.

Hofmann, S. G., & Smits, J. A. J. (2008). Cognitive-behavioral therapy for adult anxiety disorders: a meta-analysis of randomized placebo-controlled trials. *The Journal of Clinical Psychiatry, 69*(4), 621–632. https://doi.org/10.4088/jcp.v69n0415.

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.

March, J. S., & Mulle, K. (1998). *OCD in children and adolescents: a cognitive-behavioral treatment manual.* Guilford Press.

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review, 15*(1), 28–50. https://doi.org/10.1177/1088868310366253.

McLeod, B. D., & Weisz, J. R. (2010). The therapy process observational coding system for child psychotherapy-strategies scale. *Journal of Clinical Child and Adolescent Psychology, 39*(3), 436–443. https://doi.org/10.1080/15374411003691750.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: development, measurement, and validation. *The American Journal of Evaluation, 24*(3), 315–340. https://doi.org/10.1177/109821400302400303.

Pediatric OCD Treatment Study (POTS) Team (2004). Cognitive-behavior therapy, sertraline, and their combination for children and adolescents with obsessive-compulsive disorder: the Pediatric OCD Treatment Study (POTS) randomized controlled trial. *JAMA, 292*(16), 1969–1976. https://doi.org/10.1001/jama.292.16.1969.

Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology, 75*(6), 829–841. https://doi.org/10.1037/0022-006X.75.6.829.

Practice parameter for the assessment and treatment of children and adolescents with obsessive-compulsive disorder (2012). *Journal of the American Academy of Child and Adolescent Psychiatry, 51*(1), 98–113. https://doi.org/10.1016/j.jaac.2011.09.019.

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

Rosenthal, R., & Rosnow, R. L. (1984). Applying Hamlet's question to the ethical conduct of research: a conceptual addendum. *The American Psychologist, 39*(5), 561–563. https://doi.org/10.1037/0003-066X.39.5.561.

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.

Stewart, E., Frank, H., Benito, K., Wellen, B., Herren, J., Skriner, L. C., & Whiteside, S. P. H. (2016). Exposure therapy practices and mechanism endorsement: a survey of specialty clinicians. *Professional Psychology: Research and Practice, 47*(4), 303–311. https://doi.org/10.1037/pro0000094.

Stirman, S. W., Gutner, C. A., Langdon, K., & Graham, J. R. (2016). Bridging the gap between research and practice in mental health service settings: an overview of developments in implementation theory and research. *Behavior Therapy, 47*(6), 920–936. https://doi.org/10.1016/j.beth.2015.12.001.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*, 1273–1296. https://doi.org/10.1007/s11165-016-9602-2.

Vargas, C., Whelan, J., Brimblecombe, J., & Allender, S. (2022). Co-creation, co-design, co-production for public health – a perspective on definition and distinctions. *Public Health Research & Practice, 32*(2). https://doi.org/10.17061/phrp3222211.

Whiteside, S. P. H., Deacon, B. J., Benito, K., & Stewart, E. (2016). Factors associated with practitioners' use of exposure therapy for childhood anxiety disorders. *Journal of Anxiety Disorders, 40*, 29–36. https://doi.org/10.1016/j.janxdis.2016.04.001.

Wolitzky-Taylor, K., Zimmermann, M., Arch, J. J., De Guzman, E., & Lagomasino, I. (2015). Has evidence-based psychosocial treatment for anxiety disorders permeated usual care in community mental health settings? *Behaviour Research and Therapy, 72*, 9–17. https://doi.org/10.1016/j.brat.2015.06.010.

Zima, B. T., Murphy, J. M., Scholle, S. H., Hoagwood, K. E., Sachdeva, R. C., Mangione-Smith, R., Woods, D., Kamin,

H. S., & Jellinek, M.. National quality measures for child mental health care: background, progress, and next steps. *Pediatrics, 131*(Suppl 1), S38–S49. https://doi.org/10.1542/peds.2012-1427e.

Zvolensky, M. J., Rogers, A. H., Mayorga, N., Shepherd, J., Bakhshaie, J., Garza, M., Viana, A. G., et al. (2022). Perceived discrimination, experiential avoidance, and mental health among Hispanic adults in primary care. *Transcultural Psychiatry, 59*(3), 337–348. https://doi.org/10.1177/13634615211038159.