Routledge
Taylor & Francis Group

Journal of Clinical Child & Adolescent Psychology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hcap20

What is the Status of Multi-Informant Treatment Fidelity Research?

Bryce D. McLeod, Nicole Porter, Aaron Hogue, Emily M. Becker-Haimes & Amanda Jensen-Doss

To cite this article: Bryce D. McLeod, Nicole Porter, Aaron Hogue, Emily M. Becker-Haimes & Amanda Jensen-Doss (2023) What is the Status of Multi-Informant Treatment Fidelity Research?, Journal of Clinical Child & Adolescent Psychology, 52:1, 74-94, DOI: 10.1080/15374416.2022.2151713

To link to this article: https://doi.org/10.1080/15374416.2022.2151713

JOURNAL OF CLINICAL CHILD & ADOLESCENT PSYCHOLOGY
2023, VOL. 52, NO. 1, 74–94
https://doi.org/10.1080/15374416.2022.2151713

Routledge
Taylor & Francis Group

EVIDENCE BASE UPDATES

What is the Status of Multi-Informant Treatment Fidelity Research?

Bryce D. McLeod [a], Nicole Porter [b], Aaron Hogue [b], Emily M. Becker-Haimes [c] and Amanda Jensen-Doss [d]

[a]Department of Psychology, Virginia Commonwealth University; [b]Partnership to End Addiction; [c]Department of Psychiatry, University of Pennsylvania; [d]Department of Psychology, University of Miami

ABSTRACT

Objective: The precise measurement of treatment fidelity (quantity and quality in the delivery of treatment strategies in an intervention) is essential for intervention development, evaluation, and implementation. Various informants are used in fidelity assessment (e.g., observers, practitioners [clinicians, teachers], clients), but these informants often do not agree on ratings. This scoping review aims to ascertain the state of science around multi-informant assessment of treatment fidelity.

Method: A literature search of articles published through December 2021 identified 673 articles. Screening reduced the number of articles to 44, and the final study set included 35 articles.

Results: There was substantial variability across studies regarding study design, how fidelity was operationalized, and how reliability was defined and assessed. Most studies evaluated the agreement between independent observers and practitioner-report, though several other informant pairs were assessed. Overall, findings suggest that concordance across fidelity informants was low to moderate, with a few key exceptions.

Conclusions: It is difficult to draw clear conclusions about the degree to which single versus multiple informant assessment is needed to produce an accurate and complete picture of treatment fidelity. The field needs to take steps to determine how to leverage multi-informant assessment to accurately assess treatment fidelity.

Successful implementation of evidence-based interventions (EBIs) in community settings (e.g., mental health centers, schools) requires fidelity procedures designed to ensure that an intervention is delivered according to specified principles, procedures, and techniques (Aarons et al., 2011; Perepletchikova, 2011). Accurate assessment of treatment fidelity (i.e., the extent to which the quantity and quality of delivery maps onto the techniques specified in an intervention protocol; Regan et al., 2019) is critical to intervention development, evaluation, and implementation (Sutherland & McLeod, 2022). Several approaches to fidelity assessment exist that rely on various informants (e.g., observers, practitioners [clinicians, teachers], supervisors; e.g., Hogue et al., 2008; McLeod, Sutherland et al., 2022), but the different fidelity informants often do not agree in their judgments (Hogue et al., 2014; McLeod, Sutherland, et al., 2022). Though not empirically established, independent observer informants (e.g., trained, expert coders) are often considered the "gold standard" (e.g., McLeod et al., 2009). However, independent observer informants typically utilize methods that are not feasible for routine practice (e.g., coding treatment session recordings). As such, the field has been increasingly engaged in efforts to increase concordance between more pragmatic methods and independent observer reports, with the goal of being able to preference more pragmatic methods for routine use (Hogue, 2022).

A converging operations approach to conceptualizing multi-informant data is the predominant approach used to develop and evaluate fidelity measures (see De Los Reyes et al., 2013). This approach assumes that various informants assess the same target behavior. Within this conceptualization, the high overlap between informants is evidence of construct validity (convergent), whereas low overlap is seen as an error. Thus, when low correspondence between fidelity informants is observed, it is seen as error to be remedied by improving fidelity measurement – for example, improving the design of a self-report measure if it does not adequately converge with observer report (McLeod, Sutherland, et al., 2022). This is a valuable pursuit if the premise of converging operations is correct. However, if the various informants provide unique and independently meaningful information about treatment fidelity, solely seeking to reduce error is counterproductive.

CONTACT Bryce D. McLeod bmcleod@vcu.edu Department of Psychology, Virginia Commonwealth University, 806 West Franklin Street, Richmond, VA 23284- 2018

Given the pattern of discrepant informant ratings in treatment fidelity research, it is helpful to consider whether a multi-informant approach may provide an alternative way to conceptualize informant discrepancies in fidelity research. A multi-informant approach assumes that informants provide unique and independently meaningful information about target constructs (De Los Reyes, 2011; De Los Reyes & Kazdin, 2005; De Los Reyes et al., 2009, 2013). This viewpoint would suggest that efforts to advance fidelity measurement may need to focus on using multi-informant ratings to characterize treatment fidelity accurately. As a first step toward understanding whether converging operations or a multi-informant framework should be applied to fidelity assessment, this paper will ascertain the state of the field in terms of what we know about the degree to which different informants of treatment fidelity agree with one another. To accomplish this goal, we first define treatment fidelity and then review the different informants used for treatment fidelity, highlighting the unique perspectives each informant may bring to fidelity assessment. Then, we present a scoping review focused on empirical papers that evaluate the correspondence between two or more informants of treatment fidelity. Findings are discussed in terms of what the empirical work says about single- versus multi-informant treatment fidelity assessment. We conclude by presenting a research agenda to advance the goal of accurately assessing treatment fidelity.

Treatment Fidelity

Several terms and definitions have been used to characterize treatment fidelity, including treatment integrity, intervention integrity, implementation fidelity, and treatment adherence (McLeod et al., 2009; Perepletchikova & Kazdin, 2004, 2005; Sanetti et al., 2020). Here we use the term treatment fidelity, which represents a rather narrow definition that is focused on two components related to the delivery of an EBI (McLeod, Southam-Gerow et al., 2013): (a) Quantity, defined as adherence to and extensiveness of specific treatment techniques and attention to relevant treatment themes; and (b) Quality, defined as competence (skillfulness; responsiveness) in delivery of specific treatment techniques (see Table 1 for examples of how quantity and quality have been operationalized and measured to date). Measuring these components in different ways can support efforts to develop, evaluate, and implement EBIs.

Informants of Treatment Fidelity

Table 2 presents an overview of the sources of information each fidelity informant relies on, the possible biases they bring to fidelity assessment, candidate correlates of accurate fidelity ratings, and likely sources of measurement error.

Independent Observers

Independent observers are trained raters who evaluate fidelity based on the coding of data sources to capture what is happening in treatment. Observational coding of sessions by independent raters has long been considered the gold-standard method of EBI fidelity assessment (Hogue et al., 1996), providing non-participant evaluation as to whether practitioners deliver an EBI according to the letter (i.e., adherence) and spirit (i.e., competence) of the protocol with a given client set. This method entails training raters to reliably recognize a roster of techniques and then code recorded or live sessions; typically, raters are naïve to practitioner and case identities, and sessions are randomly selected (McLeod et al., 2013). This method provides a high level of rigor in fidelity assessment, yielding data that are relatively objective, procedurally rich, and ecologically valid about how practitioners implement EBIs in both controlled research (e.g., Hogue et al., 2014; Martino et al., 2009) and naturalistic field settings (e.g., Hurlburt et al., 2010; McLeod, Martinez, et al., 2022). At the same time, this method is resource-intensive, requiring strong trainer expertise to train raters to achieve sufficient interrater reliability and conduct coding while simultaneously boosting rater training to prevent reliability degradation (Hill, 1991). This method also focuses on session recordings or observations as its data source, so it may miss practitioner behaviors that do not occur on recordings (e.g., reviewing feedback data outside of sessions to support clinical decision-making in the case of measurement-based care; Jensen-Doss et al., 2020). Sometimes, sessions are audio-recorded rather than video-recorded, which may make coding the responsiveness aspects of treatment quality more difficult.

Another data source that independent observers can use is behavioral rehearsals that role-play simulations between a practitioner and actor (Beidas et al., 2016). This has been deemed a cost-effective alternative to coding sessions that can feasibly enhance EBI training efforts and validly capture EBI fidelity in an analog fashion (Beidas et al., 2014). Early returns are promising, with one study finding that behavioral rehearsal coding performed comparably to session coding – besting both practitioner-report and chart-stimulated recall methods – across various indices of EBI adherence (Becker-Haimes et al., 2022). Still, this method has generated modest practitioner participation rates outside research settings (e.g., Dorsey et al., 2017). It cannot yield data on actual EBI delivery to real-world cases by design.

Table 1. Description and examples of quality and quantity fidelity indicators.

Terms and definitions	Example anchors and scales
Quantity	
Thoroughness/Intensity: depth, complexity, or persistence with which a practice is delivered	• Child Therapy Process Rating System (CTPRS): 5-point scale from 1 = *pursued fleetingly* to 5 = *pursued intensely*; (Hurlburt et al., 2010)
Frequency: number of times a given practice is delivered or count of the number of different practices delivered	• 7-point scale of 1 = *not at all*, 2 = *a little (once)*, 3 = *infrequently (twice)*, 4 = *somewhat (3–4 times)*, 5= *quite a bit (5–6 times)*, 6 = *considerably (6 times/more depth in interventions)*, 7 = *extensively (high frequency/characterizes entire session)*; (Martino et al., 2009)
Extensiveness: thoroughness plus frequency with which a practice is delivered; provides an estimate of *dosage*	• Core Elements of Family Therapy (CEFT; Hogue et al., 2019): 5-point scale of 0 = *not at all*, 1 = a *little bit*, 2 = *moderately*, 3 = *quite a bit*, 4 = *extensively*; (Hogue et al., 2021) • Therapy Process Observational Coding System for Child Psychotherapy – Revised Strategies scale (TPOCS-RS; McLeod et al., 2015): 7-point scale of 1 = *not at all*, 4 = *considerably*, 7 = *extensively*; (McLeod, Martinez, et al., 2022)
Occurrence: whether the practice is delivered or not	• Dichotomous item of 0 = *No*, 1 = *Yes*; (Breitenstein et al., 2010; Brookman-Frazee et al., 2020); (Herschell et al., 2020)
Duration: length of time practice is delivered	• Minutes; (Gumport et al., 2019)
Quality	
Skillfulness: expertise, flexibility, mastery	• Cognitive Therapy Scale – Revised (CTS-R; Blackburn et al., 1997, 2001): 7-point scale of 0 = *incompetent* to 6 = *expert*; (Beale et al., 2020; Loades & Myles, 2016; Mathieson et al., 2009) • Independent Tape Rater scale (ITRS; Martino et al., 2008): 7-point scale from 1 = *very poor* to 7 = *excellent*; (Martino et al., 2009)
Appropriateness: suitability of practice and timing for client	• Adherence/Competence Scale for Supportive Expressive Counseling for Cocaine Dependence (ACS-SEC; Barber et al., 1996): 7-point scale from 1 = *low* to 7 = *high*; (Dennhag et al., 2012)
Comprehension: client understanding of intervention/content	• 3-point scale of 1= *no comprehension*, 2 = *partial comprehension*, 3 = *full comprehension*; (Ward et al., 2013)
Competence	• Treatment Integrity Measure for Early Childhood Classrooms (TIMECS; Sutherland & McLeod, 2015a, 2015b): skillfulness, clarity, responsiveness on 7-point scale of 1 = *very poor*, 3 = *acceptable*, 5 = *good*, 7 = *excellent*; (McLeod, Sutherland et al., 2022)
Quality/quantity combined	
Extensiveness and skill	• Pivotal Response Training Fidelity assessment tool: 3-point scale of 1 = *component not used correctly*, 2 = *component used correctly but not enough*, 3 = *component consistently used and correctly (mastery)*; (Dickson & Suhrheinrich, 2021)
Frequency and skill	• Motivational Interviewing Treatment Integrity (MITI; Moyers et al., 2010): 5-point scale from 1 = *low* to 5 = *high*; (Mullin et al., 2016) • Fidelity Checklist Competence Scale: 3-point scale of 1 = *skill rarely or never demonstrated*, 2 = *skill emerging, needs further development*, or 3 = *skill demonstrated and done well*; (Breitenstein et al., 2010)

Note: Where appropriate, original scale citations are provided along with the citation of the reviewed study.

Because of this latter limitation, this method might be better conceptualized as assessing practitioner capacity to deliver an intervention rather than their level of fidelity.

Independent observers can also code practitioner work products, such as case notes (e.g., Jensen-Doss et al., 2008) or worksheets completed as part of an EBI (e.g., Stirman et al., 2021). This method has the advantage of being more feasible than direct observation or behavioral rehearsals. Still, its validity is contingent upon the quality and scope of the data sources available for coding. For example, case notes have the advantage of being routinely collected as part of care. Still, practitioners may vary in how much information they provide about the content of sessions. The information they provide may not be sufficient to assess issues such as how competently interventions are administered. Completing worksheets is an integral part of many EBIs. Coding them can determine whether they were completed and with what level of skill (e.g., Stirman et al., 2021) but cannot directly capture how well practitioners explained the worksheets or other essential aspects of treatment.

Vested Observers

In contrast to independent ratings, sources of vested observational fidelity data include individuals such as field supervisors (e.g., Dickson & Suhrheinrich, 2021) or expert consultants (e.g., Peavy et al., 2014) judging the performance of practitioners they are training or supervising. These informants are compromised in their objectivity to the degree that they are knowledgeable about (and vested in) both the practitioners and cases being assessed, though arguably, there are unique benefits to using observational raters with expertise in the given EBI or target cases (e.g., fidelity-focused consultation; Caron & Dozier, 2021). The objectivity of vested observers is further eroded when they judge fidelity based on "second hand" data sources, such as case reporting by trained practitioners (e.g., Ward et al., 2013).

Table 2. Key considerations to guide interpretation of fidelity informants.

Informant	Sources of information	Biases	Correlates	Sources of measurement error
Independent observers	Session video or audio recordings or live observation of ongoing sessions Behavioral Rehearsals Work Products	May lack formal clinical training, especially in the research context	Level of training in EBI Level of training in coding system anchors	Behavior not caught on tape Behavior not reported in work products Methodological challenges associated with the coding system
Vested Observers	Session video or audio recordings or live observation of ongoing sessions Behavioral Rehearsals Work Products AND practitioner report in supervision/ consultation	Social desirability Halo effects	Level of training in EBI Level of training in coding system anchors	Behavior not caught on tape Behavior not reported in work products Methodological challenges associated with the coding system Information filtered through practitioners in supervision
Practitioner self-report	Own self-reflection of behavior	Social desirability Biased based on how they felt clients experienced it	Level of training in and attitude toward EBI Level of training in coding system anchors Complexity or emotional nature of the session delivered Timing of self-rating Practitioner self-efficacy with EBI	Methodological challenges associated with the fidelity tool.
Client self-report	Observation of/experience with their therapist in session	Social desirability Biased based on what they learned/remember rather than what actually happened.	Possible interrelatedness of fidelity (particularly competence) and their ability to report on fidelity (if intervention is delivered poorly, it may be harder for them to rate) Level of training in coding system anchors	Methodological challenges associated with the fidelity tool

Practitioner-Report

Another group of informants are practitioners (e.g., clinicians, teachers) self-reporting the services they have delivered. For example, practitioners have completed post-session treatment fidelity checklists evaluating the treatment provided (e.g., Chapman et al., 2008), just as school teachers have reported on their delivery of classroom-based interventions designed to prevent the development of emotional and behavioral disorders (e.g., Gresham et al., 2017; McLeod, Sutherland, et al., 2022). Practitioner self-report may be a more sustainable and less burdensome approach to understanding treatment fidelity, given the ease of assessment completion (McLeod, Sutherland, et al., 2022). However, research on how practitioners can reliably and accurately report their intervention delivery is mixed. Practitioners are not typically trained to code fidelity and may naturally focus more on client reactions to techniques than on quantifying their behavior. They also may be subject to a social desirability bias if they report on behaviors that practitioners know they are expected to engage in. By the same token, practitioners have access to case conceptualization and can report on techniques delivered outside a treatment session that are inaccessible by an independent observer (Hogue et al., 2014).

Client-Report

A final group of informants are individuals reporting on the services they receive. For example, clients and caregivers in behavioral care have completed post-session checklists regarding the strategies their practitioners used (Chaffin et al., 2016; Henggeler et al., 2002). One potential limitation of these measures is that clients and caregivers may lack the knowledge to accurately report on the delivery of techniques, particularly the quality with which they were delivered. They may also feel pressure to report that their practitioners are doing techniques if they are concerned that indicating otherwise would have negative consequences for their practitioners. Still, to the extent that measures can accurately describe the techniques, these measures have been proposed as a way to go beyond assessing whether a technique occurred and toward evaluating the degree to which clients and caregivers "got" that the strategy occurred, providing evidence of technique impact (Chaffin et al., 2016; Henggeler et al., 2002).

Many of these informants have been compared to one another in various combinations using various analytic methods. Below, we define the terminology used in prior multi-informant treatment fidelity research and then provide descriptions and frequencies of different study design components, including treatment fidelity targets,

context, informant comparison pairs, and analytic approaches. To draw conclusions about the extent to which different informants agree on treatment fidelity ratings, we organize the findings of our review by informant comparison pair. Because most of the prior work compares practitioner self-report ratings to observational ratings (both vested and independent observers), we start with a review of those studies before presenting research comparing practitioner self-report ratings to client-report ratings and then comparing pairs of informants who do not include practitioners. For all comparison pairs, we note the analytic approach and separately describe research comparing quantity and quality indicators of fidelity.

Method

Our literature search focused on articles published through December 2021 that examined concordance between fidelity informants. The search was conducted in PsycInfo and EBSCO and used the following search terms: (Therapist OR therapy OR psychotherapist OR psychotherapy AND adherence OR adhering OR adherent OR adhere OR competence OR competent OR competency OR competencies OR integrity OR fidelity AND informant OR discrepanci*). A total of 673 articles were identified and screened, 44 were flagged for a full review. The 44 articles identified as potentially eligible for inclusion were reviewed by two members of the authorship team (MJD, BDM), who came to a consensus on the final set of 35 articles included for review based on the following criteria: (a) Study looked at concordance between two treatment fidelity informants, and (b) No single case experimental designs. Two additional members of the authorship team (NP, AH) extracted information about how each study defined and operationalized fidelity, which informants were examined, the setting of the study, and the primary findings related to informant agreement reported.

Results

Table 3 presents an overview of the basic design for each study included for review. Overall, there was substantial heterogeneity across studies, beginning with how authors defined and operationalized "fidelity" as a construct. Results were organized into three categories of fidelity ratings: quantity, quality, and both. The designation of "both" was reserved for studies that assessed quantity and quality with separate rating scales. Some studies evaluating quality utilized a measure that included a quantity judgment (e.g., Dickson & Suhrheinrich, 2021), and we designated these as quality studies rather than stipulating a fourth "hybrid" category. Seventeen studies captured ratings of quantity, 11 captured quality ratings, and seven captured both fidelity components. There were other differences in how authors defined and operationalized each component. Among studies addressing quantity, operationalization consisted of indexing the presence or absence of a specific intervention, how extensively an intervention was delivered, the duration of the intervention, the number of intervention strategies delivered, and the dose (variously defined) of intervention. The authors used terms such as adherence, fidelity, and integrity. Studies on quality comprised a mix of focus on competency, skillfulness of delivery, practitioner confidence, appropriateness of the intervention, and client comprehension of the intervention delivered. Targets of fidelity assessments also varied, with some studies examining fidelity of a formal manualized intervention ($n = 10$; e.g., Alternatives for Families: A Cognitive Behavioral Therapy, Herschell et al., 2020); others focused on indexing fidelity to an evidence-based practice broadly ($n = 24$; e.g., cognitive behavioral therapy, Motivational Interviewing, see Brosan et al., 2008 and Mullin et al., 2016 respectively), and two focused on measuring practices within treatment (or business) as usual (see Hurlburt et al., 2010; McLeod, Sutherland, et al., 2022). Most studies examined practitioner fidelity within the context of traditional mental health settings ($n = 22$; e.g., community or other outpatient mental health settings, see e.g., Martino et al., 2009), although some studies examined fidelity in different settings ($n = 7$; e.g., schools, home care, see e.g., McLeod, Sutherland, et al., 2022 and Caron & Dozier, 2021 respectively) or in multiple settings ($n = 4$; e.g., Couturier et al., 2021).

Concerning informant concordance, studies examined a range of informants. All studies but four focused on examining concordance between practitioner-reported fidelity and fidelity rated by an observer. The largest proportion of studies ($n = 26$) compared practitioners to independent observers (e.g., unbiased experts or research staff members without direct case involvement; e.g., Caron et al., 2019). A smaller number focused on examining concordance between practitioners and a vested observer ($n = 9$; i.e., an expert consultant or supervisor working with the practitioner, see e.g., McManus et al., 2012) or client-report ($n = 5$; e.g., client or caregiver reported fidelity, see e.g., Chapman et al., 2013). A subset of studies ($n = 16$) focused on comparing fidelity ratings between two observers (e.g., expert consultant ratings vs. supervisor ratings, research staff ratings vs. caregiver; see e.g. Couturier et al., 2021). Many studies examined multiple comparison pairs ($n = 12$).

Table 3. Study design.

	Comparison pair[a]				Fidelity construct		Treatment characteristics	
	Provider SR vs independent observer	Provider SR vs vested observer	Provider SR vs other participant self-report	Other vs other	Type of coder judgment	Author term for coder judgment	Model[b]	Setting[c]
Beale et al. (2020)	E				Quality	Competence	CBT	CMHC
Breitenstein et al. (2010)	RA				Both	Adherence, Competence	PP	CBDC
Brookman-Frazee et al. (2020)	RA				Quantity	Occurrence, Extensiveness	Multiple	CMHC
Brosan et al. (2008)	E				Quality	Competence	CBT	CMHC
Caron et al. (2019)	RA, S			S v. RA	Quality	Competence, Confidence	CBT, TAU	SBMHC
Caron and Dozier (2021)		E			Quality	Fidelity, Adherence, Competence	ABC	HB
Carroll et al. (1998)	RA				Quantity	Adherence (Extensiveness)	CBT	AC
Chapman et al. (2013)		E	Y, CG	Y, CG, RA v. E	Quantity	Adherence (Occurrence)	CM	CMHC
Chapman et al. (2008)			Y, CG		Quantity	Adherence	CM	Mixed
Chevron and Rounsaville (1983)	E	S		E v. S	Quality	Skillfulness	IPT	CMHC
Couturier et al. (2021)		E		E v. CG, E v. RA	Quantity	Fidelity (Adherence)	FBT	Mixed
Dennhag et al. (2012)				E v. S	Both	Competence, Adherence	SE, CT, IDC	Not reported
Dickson and Suhrheinrich (2021)	RA	S		RA v. S	Quality	Fidelity	PRT	CMHC, HB
Gresham et al. (2017)	RA			Work prod	Quantity	Integrity	GBG	C
Gumport et al. (2019)	RA				Quantity	Fidelity (duration)	Sleep/ Circadian	CMHC
Herschell et al. (2020)	E		CG	E v. CG	Quantity	Adherence	AF-CBT	CMHC
Hogue et al. (2014)	RA				Quantity	Fidelity (target, focus)	MDFP	CMHC
Hogue et al. (2015)	RA				Quantity	Adherence, Fidelity (Extensiveness)	FT, MI/CBT	CMHC
Hogue et al. (2021)	RA				Quantity	Fidelity (dose)	FT. MI	CMHC
Hurlburt et al. (2010)	RA				Quantity	Strategies Pursued	TAU	CMHC
Koddebusch and Herrmann (2019)	RA		A	RA v. A	Quality	Competence	CBT	UBC
Loades and Myles (2016)	E				Quality	Competence	CBT	CMHC
Martino et al. (2009)	RA, S				Both	Adherence, Competence	MET, TAU	CMHC
Masia Warner et al. (2013)	E				Both	Adherence (occurrence), Competence (skill, appropriateness)	CBT	SBMHC
Mathieson et al. (2009)	E	S		S v. E	Quality	Competence	CBT	CMHC
McGrew et al. (2011)				S (on site) v. E (phone)	Quantity	Fidelity	ACT	CMHC
McGrew et al. (2013)				RA v. E (phone)	Quantity	Fidelity	ACT	CMHC
McLeod et al. (2021)	RA				Both	Adherence (dose), Competence	CBM	C
McManus et al. (2012)		S			Quality	Competence	CBT	Not reported
Mullin et al. (2016)	RA				Quantity	Fidelity: skill	MI	Mixed

(*Continued*)

Table 3. (Continued).

	Comparison pair[a]				Fidelity construct		Treatment characteristics	
	Provider SR vs independent observer	Provider SR vs vested observer	Provider SR vs other participant self-report	Other vs other	Type of coder judgment	Author term for coder judgment	Model[b]	Setting[c]
Peavy et al. (2014)	RA	S, E		S v. E	Both	Adherence (extensiveness), Competence	TSF	CMHC
Rollins et al. (2016)				E (on-site) v. E (by phone) v. E	Quantity	Fidelity (program/agency characteristics: present vs not present)	ACT	VAMHC
Rozek et al. (2018)	E/RA	S	A	S v. E/RA	Quality	Competence	CBT	CMHC
Sheshko et al. (2020)	RA				Quantity	Adherence	PP	CMHC
Ward et al. (2013)	RA			RA v Consultation records	Both	Coverage, Practice, Client comprehension	MATCH, TAU	CMHC

Note: [a]A = Adult client, CG = Caregiver, E = Expert/Consultant, RA = Research Assistant, S = Supervisor, SR = Self-report, Y = Youth client.
[b]ABC = Attachment and Biobehavioral Catch-up, ACT = Assertive Community Treatment, AF-CBT = Alternatives for Families: A Cognitive Behavioral Therapy, CBM = Classroom Behavioral Management, CBT = Cognitive Behavior Therapy, CM = Contingency Management, CT = Cognitive Therapy, FT = Family Therapy, FBT = Family-based Treatment, GBG = Good Behavior Game, IDC = Individual Drug Counseling, IPT = Interpersonal Therapy, MATCH = Modular Approach to Therapy for Children, MDFP = Multidimensional Family Prevention, MET = Motivational Enhancement Therapy, PP = Parenting Program, PRT = Pivotal Response Training, SE = Supportive-expressive Therapy, TAU = Treatment as usual/Usual care, TSF = 12-Step Facilitation.
[c]AC = Ambulatory care, C = Classroom, CBDC = Community-based Day Care, CMHC = Community mental health clinic, HB = Home-based, Mixed = Combined samples or multi-site, SBMHC = School-based mental health clinic, UBC = University-based clinic, VAMHICM = Veteran Affairs Mental Health Intensive Case Management.

The overall synthesis of findings is complicated by the various metrics used across studies to index informant concordance. Sixteen studies assessed interrater reliability to understand consistency in rank order scores (e.g., Pearson's correlations, discrepancy, or difference scores; see, e.g., Brosan et al., 2008). This contrasts with studies testing interrater agreement ($n = 11$) or the absolute consensus in scores furnished by multiple judges (e.g., percent agreement, kappa; see e.g., Loades & Myles, 2016; Hurlburt et al., 2010 respectively). Fifteen studies employed methods that assessed a combination of reliability and agreement (e.g., ICCs). Other studies examined mean comparisons between separate fidelity scores (e.g., chi-squared or t-tests, $n = 18$; see e.g., Hogue et al., 2021). Some studies ($n = 22$) used a combination of analytic methods. Table 4 presents an overview of the methodologies employed by the studies.

Summary of Findings

Table 5 presents the significant findings concerning informant concordance from each study included in the review. While variations in methodology across studies preclude firm conclusions, a summary of observed patterns of concordance between informants is presented below. Note that the term "concordance" is used throughout to describe results. We use the following sets of accepted guidelines for classifying various concordance coefficients, also reporting values of individual studies where appropriate: (a) Cohen (1988) criteria for Pearson's r and non-parametric alternatives: 0 to .10 is no relation, .11 to .30 is weak, .31 to .50 is moderate, and .50 to 1.0 is strong; (b) ICC magnitudes are interpreted based on Cicchetti's (1994) criteria, which are ubiquitous in observational coding research on behavioral interventions: below .40 is poor, .40–.59 is fair, .60–.74 is good, and .75–1.0 is excellent, but it is also possible to interpret ICC magnitudes using Koo and Li's (2016) criteria recommended for behavioral measurement theory more broadly: below .50 is poor, .50–.74 is fair, .75–.90 is good, and .91–1.0 is excellent; and (c) (LeBreton & Senter, 2008) criteria for classifying r_{wg}/a_{wg} coefficient magnitude: < .30 is lack of concordance, .31–.50 weak, .51–.70 moderate, .71–.90 strong, and .91–1.0 very strong. There are no universally accepted guidelines for classifying percent agreement, so values from individual studies are reported.

Concordance Between Practitioners and Independent Observers

Of the 26 studies that examined the concordance of fidelity between practitioner-report and independent observers, 11 focused on quantity, nine focused on quality, and six focused on both. Overall, studies in this category examining quantity found practitioners tended to report higher fidelity than what was rated by independent observers, with variable concordance between informants across studies. Studies assessing concordance via Pearson's r correlations found a zero to weak correlation at the scale level ($r = .24$; Gresham et al.,

Table 4. Study analytics.

	Reliability[a]		Interrater agreement[b]			Reliability/ agreement combined[c]	Mean comparison				
	Pearson's *r*	Discrep score	% Agree.	Kappa	r_{wg} or a_{wg} index	ICC index	Chi-squared	*t*-test	ANOVA	Mixed effects modeling	Misc.
Beale et al. (2020)	X		X				X[d]				
Breitenstein et al. (2010)			X			X					
Brookman-Frazee et al. (2020)						X					Q correlations
Brosan et al. (2008)	X[d]	X						X			
Caron et al. (2019)	X	X						X			
Caron and Dozier (2021)[f]						X			X		
Carroll et al. (1998)			X	X							
Chapman et al. (2008)											Many-Facet Rasch Models
Chapman et al. (2013)											Many-Facet Rasch Models
Chevron and Rounsaville (1983)	X							X			
Couturier et al. (2021)						X			X		
Dennhag et al. (2012)	X[e]									X[e]	G Theory
Dickson and Suhrheinrich (2021)			X						X		
Gresham et al. (2017)											G Theory
Gumport et al. (2019)	X[d]										
Herschell et al. (2020)				X							
Hogue et al. (2014)						X		X			
Hogue et al. (2015)						X			X[e]		r-z transformation for ICC comparisons
Hogue et al. (2021)						X		X			
Hurlburt et al. (2010)	X			X		X		X			
Koddebusch and Herrmann (2019)	X										
Loades and Myles (2016)[f]		X	X				X	X[d]	X		
Martino et al. (2009)			X			X			X		G Theory
Masia Warner et al. (2013)			X								
Mathieson et al. (2009)[f]	X										
McGrew et al. (2011)		X				X					
McGrew et al. (2013)		X				X					Sensitivity & Specificity
McLeod et al. (2021)	X				X					X[e]	Standard mean diffs
McManus et al. (2012)	X							X			
Mullin et al. (2016)[f]	X							X			
Peavy et al. (2014)						X[e]				X[e]	
Rollins et al. (2016)		X	X			X				X	Tukey's pairwise comparison
Rozek et al. (2018)									X		
Sheshko et al. (2020)						X					
Ward et al. (2013)						X					

Note: [a]Interrater reliability: Used to address whether judges rank order targets in a manner that is consistent with other judges (LeBreton & Senter, 2008).
[b]Interrater agreement: Refers to the absolute consensus in scores furnished by multiple judges (LeBreton & Senter, 2008).
[c]Interrater reliability and agreement combined: Assesses both reliability and agreement collectively (LeBreton & Senter, 2008).
[d]Study used non-parametric alternative.
[e]Analyses accounted for nested data.
[f]Study included additional analyses testing for change in multi-informant agreement over time.

2017) and item level (50% ρ < .30, Gumport et al., 2019; 95% r < .08, Mullin et al., 2016). Studies using ICCs found poor concordance for nearly all items (ICC range =.08 to .40, Sheshko et al., 2020; ICC range =.33 to .34; Martino et al., 2009) or a significant portion of items (50% ICC <.40, Brookman-Frazee et al., 2020), except for one study finding fair to excellent concordance (ICC range =.42 to 1.00, Ward et al., 2013). Other studies utilized Cohen's kappa to assess concordance, finding only slight or fair concordance for most items (60%, Carroll et al., 1998; 89%, Herschell et al., 2020; 80%, Hurlburt et al., 2010). This is similar to one study using a_{wg} and $r_{wg,}$ that found no or weak agreement for 88% of items (McLeod, Sutherland, et al., 2022). Finally, one study that utilized mean comparison methods reported significant differences between coders (Peavy

Table 5. Summary of study findings.

	Aims	Results
Beale et al. (2020)	"This study investigated the relationship between self- and expert-rated competence – assessed via therapy recordings rated on the Cognitive Therapy Scale Revised (CTS-R) scale – for a large sample of IAPT CBT trainees during training and, for the first time, at post-training follow-up." (p. 1)	"There were positive relationships (r = .27 to .56) between self and expert CTS-R scores at all time points. The proportion of tapes demonstrating significant agreement between self and expert ratings (CTS-R difference <5 points) increased significantly across training and remained stable at follow-up." (p. 1)
Breitenstein et al. (2010)	"The objective of the study was to examine the reliability and validity of the Fidelity Checklist, a measure designed to assess group leader adherence and competence delivering a parent training intervention (the Chicago Parent Program) in child care centers serving low-income families." (p. 158)	"Agreement between group leader self-report and independent ratings on the Adherence Scale was 85%; disagreements were more frequently due to positive bias in group leader self-report. " (p. 158)
Brookman-Frazee et al. (2020)	"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." (p. 155)	"Concordance between therapist- and observer report of the extensiveness of therapist EBP strategy use was at least fair (*ICC* ≥ .40) for approximately half of the items." (p. 155)
Brosan et al. (2008)	"This study aimed to examine the accuracy of therapists' judgments about their own competence in cognitive therapy." (p. 581)	"Whilst there was a significant correlation between self-ratings and expert ratings of competence, therapists significantly over-rated their competence relative to the expert rater." (p. 581)
Caron et al. (2019)	"This study examined the concordance between clinicians,' supervisors' and independent observers' session-specific ratings of clinician competence in school-based CBT and treatment as usual (TAU)." (p. 1)	"Patterns of rater discrepancies differed between the TAU and CBT groups. Correlations with independent raters were low across groups." (p. 1)
Caron & Dozier (2021)	"The current study examined clinicians' self-coding of fidelity as a potential active ingredient of consultation for the Attachment and Biobehavioral Catch-up (ABC) intervention." (p. 237)	"Clinicians' ABC fidelity, as well as their self-coding accuracy, increased over the course of consultation. Clinicians' self-coding accuracy predicted their initial fidelity and growth in fidelity. Working alliance was also linked to fidelity and self-coding accuracy. These results suggest that clinician self-coding should be further examined as an active ingredient of consultation." (p. 237)
Carroll et al. (1998)	"In this report, we provide preliminary data on the utility of therapist self-report session checklists, modeled on the CSPRS system, as a strategy for monitoring delivery of manual-specified interventions." (p. 307)	"The poor levels of concordance between therapist and observer ratings suggest that therapist session reports may be a supplement to, but not substitute for, observer ratings." (p. 307)
Chapman et al. (2013)	"This study evaluated the accuracy of youth, caregiver, therapist, and trained raters relative to treatment experts on ratings of therapist adherence to a substance abuse treatment protocol for adolescents." (p. 674)	"Relative to treatment experts, youth and caregivers were significantly more likely to endorse the occurrence of CM components. In contrast, therapists and trained raters were much more consistent with treatment experts. In terms of practical significance, youth and caregivers each had a 97% estimated probability of indicating that a typical treatment component had occurred. By comparison, the probability was 31%, 19%, and 26% for therapists, trained raters, and treatment experts, respectively." (p. 674)
Chapman et al. (2008)	"A unique application of the Many-Facet Rasch Model (MFRM) is introduced as the preferred method for evaluating the psychometric properties of a measure of therapist adherence to Contingency Management (CM) treatment of adolescent substance use." (p. 48)	" . . . there was some indication that therapist reports might be more conservative than the reports of caregivers and/or youth." (p. 65)
Chevron and Rounsaville (1983)	"This study evaluates different methods of assessing psychotherapy skills." (p. 1129)	"Results show poor agreement among assessments of therapists' skills based on different data sources. Most important, ratings based on review of videotaped sessions were uncorrelated with those based on supervisor's discussion of process material with the therapist." (p. 1129)
Couturier et al. (2021)	"As part of a larger implementation study, we examined fidelity to Family-Based Treatment (FBT) measured by several different raters including an expert, a peer, therapists themselves, and parents, with a goal of determining a pragmatic, reliable and efficient method to capture treatment fidelity to FBT." (p. 1)	"Intraclass correlation coefficients revealed that agreement was the best between expert and peer, with excellent, good, or fair agreement in 7 of 13 items from session 1, 2 and 3. There were only four such values when comparing expert to parent agreement, and two such values comparing expert to therapist ratings. The rest of the ICC values indicated poor agreement. Scale level analysis indicated that expert fidelity ratings for phase 1 treatment sessions scores were significantly higher than the peer ratings and, that parent fidelity ratings tended to be significantly higher than the other raters across all three treatment phases. There were no significant differences between expert and therapist mean scores." (p. 1)
Dennhag et al. (2012)	"The current study examined the agreement between supervisors' and independent judges' evaluations of therapist adherence and competence in three treatments of cocaine dependence: supportive expressive therapy (SE), cognitive therapy CT), and individual drug counseling (IDC)." (p. 720)	"At the therapist level of analysis, the agreement between supervisors' and independent judges' ratings was weak for SE competence, CT adherence, and CT competence. Moderate relations were found for IDC adherence and competence. Supervisors consistently rated adherence and competence more positively than judges in CT and IDC." (p. 720)

(*Continued*)

Table 5. (Continued).

	Aims	Results
Dickson and Suhrheinrich (2021)	"Using a train-the-trainer methodology, we examine concordance between three methods of assessing fidelity (trained independent coders, supervisor evaluation and provider self-report) using a fidelity assessment tool adapted for community use." (p. 542)	"Results suggest supervisors and providers are able to use the fidelity tool, but demonstrated variable concordance, with higher concordance with trained coders for supervisors than providers." (p. 542)
Gresham et al. (2017)	"The current study used G theory to examine the dependability of direct observation, permanent products, and self-report as measures of treatment integrity when six teachers implemented the Good Behavior Game across three study sites." (p. 108)	" . . . the results indicated low correlations between the self-report assessment and the other two measures of treatment integrity; however, a significant correlation was found between permanent product and direct observation." (p. 118)
Gumport et al. (2019)	"The present study reports on the psychometric properties of the Provider-Rated TranS-C Checklist – a provider-reported fidelity measure for the Transdiagnostic Sleep and Circadian Intervention (TranS-C)." (p. 800)	"Provider-Rated TranS-C Checklist scores were positively associated with the Independent-Rater TranS-C Checklist scores demonstrating convergent validity." (p. 800)
Herschell et al. (2020)	"This study compared expert-coded behavior observations, therapist and caregiver report of therapists' adherence to nine teaching technique items assessed in treatment sessions using Alternatives for Families: A Cognitive Behavioral Therapy (AF-CBT) to determine whether other raters (outside of traditional expert coders) could effectively and accurately measure therapist adherence." (p. 92)	"Outcomes indicated strikingly different ratings across all reporters suggesting that therapist and caregiver reports may be supplement to, but not substitute for, observer ratings." (p. 92)
Hogue et al. (2014)	"This study examined therapist reliability and accuracy in rating intervention target (i.e., session participants) and focus (i.e., session content) in a manual-guided, family-based preventive intervention implemented with 50 inner-city adolescents at risk for substance use." (p. 697)	"Therapists demonstrated excellent reliability with coders for treatment targets and moderate to high reliability for treatment foci across the sample and within each phase. Also, therapists did not consistently overestimate their degree of activity with targets or foci." (p. 697)
Hogue et al. (2015)	"This study tested the reliability and accuracy of two groups of community therapists who reported on their use of family therapy (FT) and motivational interviewing/cognitive behavioral therapy (MI/CBT) interventions during routine treatment of inner-city adolescents with conduct and substance use problems." (p. 229)	"Overall therapist reliability was adequate for averaged FT ratings (ICC = .66) but almost non-existent for MI/CBT (ICC = .06); moreover, both RFT and TAU therapists were more reliable in reporting on FT than on MI/CBT. Both groups of therapists overestimated the extent to which they implemented FT and MI/CBT interventions." (p. 229)
Hogue et al. (2021)	"This study describes reliability and validity characteristics of a therapist-report measure of family therapy techniques for treating adolescent conduct and substance use problems: Inventory of Therapy Techniques for Core Elements of Family Therapy (ITT-CEFT) ." (p. 298)	"Concurrent validity analyses showed fair-to-excellent therapist reliability compared to observer ratings (ICCs range .64–.75); they showed moderate therapist accuracy compared to observer mean scores, reflecting a tendency to overestimate delivery of the techniques." (p. 298)
Hurlburt et al. (2010)	"The objective of this study is to examine the characteristics of outpatient mental health services delivered in community-based outpatient clinics, comparing information obtained from two different sources, therapists serving children and families, and observational coders viewing tapes of the same treatment sessions." (p. 230)	"Therapists reported pursuing 2.5 times more goals and strategies per session, on average, than identified by observational coders. Correspondence between therapists and coders about the occurrence of specific goals and strategies in treatment sessions was low, with 20.5% of codes having a Kappa of .4 or higher." (p. 230)
Koddebusch and Herrmann (2019)	"We developed a set of measurements assessing therapeutic competences from different perspectives: therapists' global (GloRa-T) and session self-rating (SeRa-T), clients' session rating (SeRa-C) and observer rating (CoRa-O). The psychometric properties of the measurements were investigated." (p. 15)	"When investigating the relationship between the different perspectives by analyzing the intercorrelations of the measurements, only few perspective/subscales were correlated." (p. 26)
Loades and Myles (2016)	"We aimed to explore the relationship between therapists' reflective ability and the level of agreement between self-rated competence and competence rated by an experienced CBT assessor." (p. 1)	"Trainees tended to overestimate or underestimate their competence in comparison to the independent assessors." (p. 1)
Martino et al. (2009)	"This study examined the correspondence of treatment integrity ratings (adherence and competence) among community program therapists, supervisors, and observers for therapists who used motivational enhancement therapy (MET) within a National Institute on Drug Abuse Clinical Trials Network protocol." (p. 181)	"The results suggested there was reasonable agreement between the three groups of raters about the presence or absence of several fundamental MET strategies. Moreover, relative to observers, therapists and supervisors were more positive in their evaluations of the therapists' MET adherence and competence." (p. 181)
Masia Warner et al. (2013)	"We present an initial consultation strategy to support school counselor implementation of group CBT for social anxiety and an evaluation of counselors' treatment fidelity." (p. 541)	"Counselors and consultants demonstrated good agreement for adherence, but relatively modest correspondence in competence ratings." (p. 541)
Mathieson et al. (2009)	"This paper investigates the accuracy of self-rating of competence in relation to other measures such as 'direct' assessment of videotaped sessions or supervisor ratings." (p. 43)	"Results are discussed and it is suggested that trainee self-assessment, while not found in this study to be correlated with other measure of competence, may provide important information about confidence development . . . " (p. 43)
McGrew et al. (2011)	"This study investigated the reliability and validity of a phone-administered fidelity assessment instrument based on the Dartmouth Assertive Community Treatment Scale (DACTS)." (p. 670)	"Phone and on-site assessment showed strong agreement (ICC=.87) and consensus (mean absolute difference of .07) and agreed within .1 scale point, or 2% of the scoring range, for 83% of sites and within .15 scale point for 91% of sites." (p. 670)

(*Continued*)

Table 5. (Continued).

	Aims	Results
McGrew et al. (2013)	"This study investigated the reliability and validity of a less burdensome approach: self-reported assessment." (p. 272)	"DACTS total scores obtained via self-reported assessments were reliable and valid compared with phone-administered assessment on the basis of interrater consistency (intraclass correlation) and consensus (mean rating differences). Phone administered assessments agreed with self-reported assessments within .25 scale points (out of 5 points) for 15 of 16 teams." (p. 272)
McLeod et al. (2021)	"This paper reports on the development and initial evaluation of the score reliability and validity of the Treatment Integrity Measure for Early Childhood Settings Teacher Report (TIMECS-TR), which is designed to address limitations of previous self-report treatment integrity measures that may have contributed to low correspondence with observer-rated measures." (p. 20)	"Analyses did not support the convergent score validity of the TIMECS-TR items or scale with observational ratings of the same practices. Teachers reported higher levels of practice delivery on the TIMECS-TR items relative to observer report." (p. 20)
McManus et al. (2012)	"To examine the accuracy of therapists' self-assessment of their CBT competence in relation to supervisors' assessments." (p. 292)	"There were moderate correlations between self- and supervisor assessments, and the previously reported over-estimation of CBT skills . . . was not replicated in the current sample. Instead, these groups showed *under*-estimation of their skills compared to supervisors' ratings." (p. 292)
Mullin et al. (2016)	"It also evaluates clinicians' ability to accurately self-assess their MI skills." (p. 357)	"There was little correlation between participants' self-assessment of MI skills and objective assessment." (p. 357)
Peavy et al. (2014)	"This study investigated the correspondence among four groups of raters on adherence to STAGE-12, a manualized 12-step facilitation (TSF) group and individual treatment targeting stimulant abuse." (p. 222)	"Results indicated that external raters rated most critically mean adherence – the mean of all the adherence items – and global performance. External raters also demonstrated the highest degree of reliability with the designated expert. Therapists rated their own adherence lower, on average, than did supervisors and TSF expert raters, but therapist ratings also had the poorest reliability." (p. 222)
Rollins et al. (2016)	"This study compared reliability and validity of three methods of fidelity assessment (on-site, phone-administered, and expert scored self-report) using a stratified random sample of 32 mental health intensive case management teams from the Department of Veterans Affairs." (p. 157)	"Overall, phone, and to a lesser extent, expert-scored self-report fidelity assessments compared favorably to on-site methods in inter-rater reliability and concurrent validity." (p. 157)
Rozek et al. (2018)	"This pilot study compared ratings of CBT competency from four perspectives – patient, therapist, supervisor and independent observer using the Cognitive Therapy Scale (CTS)." (p. 245)	"Analyses of variance revealed that therapist average CTS competency ratings were not different from supervisor ratings, and supervisor ratings were not different from independent observer ratings; however, therapist ratings were higher than independent observer ratings and patient ratings were higher than all other raters." (p. 245)
Sheshko et al. (2020)	"We present preliminary multimethod, multi-informant data on a new measure of adherence (*Practitioner Session Reflection Tool*) that invited practitioners to report on their adherence and content modifications to an evidence-based parenting program." (p. 290)	"The comparison of ratings across practitioner self-report and external coders yielded low correlations between informants." (p. 290)
Ward et al. (2013)	"This study sought to evaluate the agreement between therapist report and coder observation of therapy practices." (p. 44)	"Intraclass correlation coefficients (ICCs) representing coder versus therapist agreement on manual content delivered ranged from .42 to 1.0 across conditions and problem areas. Analyses revealed marked variability in agreement regarding whether behavioral rehearsals took place (ICCs from _.01 to 1.0) but strong agreement on client comprehension of therapy content and homework assignments." (p. 44)

et al., 2014). Notable exceptions to these patterns include work by Hogue et al. investigating family therapy interventions who found stronger reliability assessed via ICCs (ICC = .66, Hogue et al., 2015; ICC range =.64 to .75, Hogue et al., 2021) and found no evidence that practitioners tended to report higher quantity in one study (Hogue et al., 2014). Also, research using percent agreement reports higher concordance (e.g., 85%, Breitenstein et al., 2010; 87%, Masia Warner et al., 2013) than research using other types of concordance coefficients.

Studies in this category examining quality similarly found that in general, practitioners reported higher quality scores, with inconsistent findings for concordance between coders. Three studies using simple correlational analyses found weak correlations (r = .11 to .26, Caron et al., 2019; r = .18, Chevron & Rounsaville, 1983; r = .15, Mathieson et al., 2009), one study showed a moderate correlation (r = .48, Koddebusch & Herrmann, 2019), and one study showed a strong correlation (r = .57, Brosan et al., 2008). An additional study reported high variability across items (r = .27 to .56, Beale et al., 2020). Of note, Brosan et al., 2008, also reported significant t-test results comparing coders. Other studies using mean comparison methods similarly reported significant differences (Caron et al., 2019; Martino et al., 2009; Peavy et al., 2014; Rozek et al., 2018), yet one

study using chi-square comparison found no difference in ratings (Loades & Myles, 2016). One study using ICCs found poor to good ICCs (Ward et al., 2013). Like quantity studies, quality studies using percent agreement (versus other coefficients) found generally higher rates (average percent agreement = 58%, Dickson & Suhrheinrich, 2021; 67% Masia Warner et al., 2013). One notable exception was a study in which practitioners reported higher quality ratings (Masia Warner et al., 2013).

Concordance Between Practitioner-Report and a Vested Observer

Of the nine studies that examined fidelity concordance between practitioners and a vested observer, two focused on quantity, six focused on quality, and one focused on both. One study examining quantity used the ICC and found poor ICCs for 85% of items (ICC range = −1.33 to .35, Couturier et al., 2021). Yet, Couturier et al. (2021) found no difference in scale-level scores via ANOVA. One study using mean comparison methods found significant differences between practitioners and vested observers using *t*-tests (Peavy et al., 2014). One found differences using Many-facet Rasch Models (Chapman et al., 2013).

Of studies testing quality ratings, results were mixed. Of the studies that used Pearson's *r*, two reported weak correlations ($r = .18$, Chevron & Rounsaville, 1983; $r = .21$, Mathieson et al., 2009), and one study reported moderate correlations ($r = .35$ to .62, McManus et al., 2012). One study that used ICCs found fair to excellent ICCs for the majority of items (90%) and fair concordance at the scale level (ICC = .40, .45, Caron & Dozier, 2021). One study found no significant difference using ANOVA (Rozek et al., 2018), while another found a significant difference using *t*-tests (Peavy et al., 2014). One study utilized percent agreement, finding an average of 67% agreement across items (Dickson & Suhrheinrich, 2021).

Concordance Between Practitioner- and Client-Report

Of the five studies that examined fidelity concordance between practitioners and clients (i.e., youth and adult clients or caregivers), three focused on quantity and two focused on quality. Chapman and colleagues found youth and caregivers tended to score treatment delivery higher than practitioners using Many-facet Rasch Models (e.g., Chapman et al., 2008, 2013), and another study found only slight to fair concordance using Cohen's kappa (89% of items, Herschell et al., 2020). In terms of quality, one study found a moderate correlation using Pearson's *r* between ratings by practitioners and caregivers for only some isolated subscales (e.g., $r = .37$ for interpersonal competence; Koddebusch & Herrmann, 2019). Another study found a significant difference in ratings produced by practitioners and their adult clients using ANOVA (Rozek et al., 2018), with clients providing higher scores than their practitioners.

Concordance Between Two External Informants

Of the 16 studies that examined fidelity concordance between two external informants, seven focused on quantity, six focused on quality, and three focused on both. Overall, the results were inconsistent. In terms of quantity, two studies utilized correlations. One found a moderate correlation between direct observation and work products ($r = .43$, Gresham et al., 2017), while another found weak and large correlations between vested and independent observers based on particular treatment models ($r = .26$ and $r = .51$, Dennhag et al., 2012). Studies that utilized ICCs to compare vested and independent experts found fair to excellent results at the item level (ICC range =.47–.98, with 89% falling above .60, McGrew et al., 2013) and scale/subscale level (ICC = .69–.93, McGrew et al., 2011; ICC =.96, Rollins et al., 2016) as did studies comparing research assistants to work products (ICC range =.70 to .75, Ward et al., 2013). One study comparing caregiver report to trained peers to experts found high variability at the item level (ICC range = −0.62 to .88), with better concordance between expert coders and trained peers (50% ICC <.38) than between experts and caregivers (67% ICC <.35, Couturier et al., 2021). Another study comparing caregivers to observers found only slight agreement using Cohen's kappa (κ range =.01 to .26, 90% κ < .20, Herschell et al., 2020). Similarly, one study using Many-facet Rasch Models found variability between youth, caregiver, expert, and trained coders (Chapman et al., 2013).

Regarding quality, results were similarly mixed. Studies using Pearson's *r* found weak and moderate correlations comparing vested supervisors to independent experts ($r = .15$, Mathieson et al., 2009; and $r = .31$, Chevron & Rounsaville, 1983, respectively) and to independent observers ($r = .17$ to .21, Caron et al., 2019). Two studies found variability of coefficient magnitude across subscales or treatments, one comparing vested supervisors to independent experts ($r = .26$ and .51, Dennhag et al., 2012) and one comparing adult clients to independent observers (moderate correlation for only one scale; Koddebusch & Herrmann, 2019). On on hand, one study utilized mean comparisons and found significant differences between vested supervisors and independent observers, as well as independent experts and independent observers (Peavy et al., 2014). On the other hand, one study using percent agreement found high rates of agreement between vested supervisors and

independent observers (72% on average, Dickson & Suhrheinrich, 2021), and another using ANOVA found no significant difference between these informants (Rozek et al., 2018).

Discussion

Our scoping review revealed that it is difficult to draw clear conclusions about the degree to which single versus multiple informants are needed to provide an accurate picture of treatment fidelity. The reviewed studies appear to adopt a converging operations framework focused on indexing agreement between informants. The main limitations of the extant literature include (a) a lack of consistent terminology, making it challenging to identify studies examining similar questions; (b) different approaches and metrics used to analyze agreement data, including several studies relying on outdated or inappropriate statistics, and (c) no focus on establishing whether each informant brings unique and meaningful information to fidelity assessment. These limitations make it difficult to draw broad conclusions from this set of studies about the value of single versus multiple informant assessment. The only somewhat consistent findings that emerged were (a) practitioners generally show low to modest agreement with independent observers; and (b) practitioners generally provide significantly higher mean ratings than independent observers; however, not all studies report these patterns (e.g., Hogue et al., 2015). To advance our understanding of how best to assess fidelity, there is a need for more rigorous research that (a) assesses concordance between informants with rigorous methods and (b) moves beyond simply examining concordance to trying to understand the degree to which convergence and divergence between informants represent error versus meaningful differences in perspective. As such, the rest of the paper will focus on concrete steps the field can take to address these issues and help bring the field forward. Using De Los Reyes et al. (2013, 2022) as a guide, we will focus on the following steps: (a) the importance of terminology, (b) establishing score reliability and validity of fidelity measures, (c) ruling out methodological factors that might account for discrepancies, and (d) how to design studies to determine if specific mechanisms explain differences in informant reports.

Standardize Terminology

One key to moving the field forward is to adopt a consistent set of terms to refer to treatment fidelity. Treatment fidelity primarily focuses on the delivery of practices found within a specific intervention protocol, which is most appropriate for research questions early in the translational pipeline that focus on estimating protocol adherence/competence (i.e., following the procedures specified within a particular intervention protocol; Regan et al., 2019). Other similar fidelity terms exist, such as practice sequencing (i.e., the extent to which the order of practices prescribed by an intervention protocol is followed; Park et al., 2015) and consultant recommendations (i.e., expert recommendations regarding techniques from a specific protocol to deliver; Regan et al., 2019). Effectiveness and implementation research sometimes focus less on the delivery of specific intervention protocols and more on general practice patterns (i.e., the degree to which the quantity and quality of practices delivered map onto the research literature; Brookman-Frazee et al., 2020). Since these research questions do not deal with specific intervention protocols, the terms "quantity" and "quality" are more appropriate as barometers of fidelity (McLeod, Sutherland, et al., 2022). Thus, going forward, we propose the use of three sets of terms: those specific to intervention protocols—i.e., fidelity: adherence and competence to protocols, sequencing, and consultant recommendations (Park et al., 2015; Regan et al., 2019); those specific to general practice use—i.e., fidelity: quantity and quality of practice use (McLeod, Sutherland, et al., 2022); and those that are overarching and can be applied to both—i.e., fidelity: quantity, quality.

Are Our Measures Robust Enough to Yield Trustworthy Findings?

A second key to moving the field forward is demonstrating that measures from each informant evidence score reliability and validity. As noted earlier, observational assessment of treatment fidelity is often considered a gold-standard assessment (McLeod et al., 2009), yet empirical evidence does not support this assertion. A traditional definition of a gold standard treatment fidelity measure might be one that demonstrates sufficient reliability (precision) and validity (accuracy) to be a useful method to support research (e.g., manipulation check) or clinical (e.g., quality improvement) applications (Kraemer et al., 2003). However, we argue that feasibility of implementation should be part of the definition, as a fidelity measure that meets reliability and validity standards that cannot be implemented in certain contexts due to cost and time constraints lacks utility. We are unaware of any treatment fidelity measures that have met these reliability, validity, or feasibility standards and wonder if it is reasonable to expect that any single method could achieve all of these standards. Some exemplar observational measures stand out as

demonstrating reliability and some forms of validity (e.g., Inventory of Therapy Techniques for Core Elements of Family Therapy; Hogue, 2022); however, convergent validity between fidelity measures that utilize the identical method (observational coding) to assess the same fidelity component (adherence to a specific intervention protocol) is rarely demonstrated, which raises important questions about construct validity (i.e., do these measures assess adherence or competence).

Critically, our scoping review indicated little consistency in the terms and metrics used to evaluate reliability. Reliability and agreement are often used interchangeably to describe concordance between ratings produced by different informants using various analytic techniques. Reliability and agreement, although related, are distinct facets of coder concordance with other indices for measurement (see Gisev et al., 2013; LeBreton & Senter, 2008). This nuance is frequently overlooked in the literature. Interrater agreement (IRA; aka accuracy) refers to absolute consensus in scores or the degree to which scores are identical. IRA is most commonly indexed by mean comparisons, percent agreement, and r_{wg}/a_{wg} coefficients. In contrast, interrater reliability (IRR) refers to the relative consistency in ratings or rank order. IRR is often measured via Pearson correlations. Different still are indices designed to measure IRA + IRR combined, that is, the consistency in ratings as a function of absolute consensus. The most commonly used index of IRA+IRR is the intraclass correlation coefficient (ICC). Greater attention to the distinctions, advantages, and disadvantages of assessment methods of IRA, IRR, and IRA + IRR, and their corresponding interpretation, is required to progress the field forward.

It is recommended here and elsewhere (e.g., Chaturvedi & Shweta, 2015; Goodwin, 2001) that researchers estimate and report informant concordance in multiple ways to best understand both the equivalence of scores and the consistency of scores across informants. To assess IRA, researchers are encouraged to use Cohen's kappa (Cohen, 1960) as an alternative to simple percent agreement (Hallgren, 2012). Cohen's kappa and its derivatives (e.g., weighted kappa) take into account agreement by chance and are, therefore, a more precise estimate of agreement. IRR can be measured via Pearson correlation. Although familiar and relatively easy to interpret, Pearson correlation coefficients do not consider systematic differences in informants' use of the rating scale (e.g., their respective means). Therefore, paired t-tests or repeated measures ANOVA are recommended to be used as an adjunct to estimate reliability (Hartmann, 1977). Informant discrepancy scores have also been utilized as an intuitive index of IRR. Discrepancy scores can be useful descriptively to understand the direction of over/under scoring between informants. However, researchers should proceed with caution when using discrepancy scores to predict outcome variables due to mathematical and statistical limitations related to information redundancy (see Laird, 2020). The gold standard for measuring IRA + IRR is the ICC, the most widely utilized statistic in this review to assess informant concordance. One disadvantage of the ICC is low variation in ratings, and small sample sizes can cause artificially low and misleading ICC values. It is recommended that percent agreement be used descriptively in the case of low ICCs to diagnose if and when low variation contributes to unreliable ICCs (Chaturvedi & Shweta, 2015). More complex methods of assessing informant concordance are emerging, including generalizability theory (e.g., Dennhag et al., 2012; Gresham et al., 2017) and Many-facet Rasch Modeling (e.g., Chapman et al., 2008, 2013). Although these methods may provide more information about informant agreement and reliability, analytic complexity has been a limiting factor in their widespread adoption.

Can We Rule Out Methodological Factors?

As noted above, conventional wisdom in the fidelity research field suggests that discrepancies between informants represent error, with independent observers considered the "gold standard" informant. As detailed in Table 2, several potential sources of measurement error could decrease agreement between informants. However, the field of fidelity research has only begun to grapple with whether methodological factors could potentially account for discrepancies. The larger informant agreement literature points to several important future directions to further our understanding in this area. First, this literature suggests that more concrete, behavioral items (e.g., my child attends school every day/I attend school every day) would have higher overall agreement than less observable constructs (e.g., my child worries every day/I worry every day; Comer & Kendall, 2004). There are clear parallels to fidelity measurement as the fidelity research field has not done a good job of defining what level of specificity should be used for items (McLeod et al., 2013). Items on extant measures vary widely in their degree of specificity (McLeod et al., 2013; Schoenwald, 2011). For example, a rating item asking for judgment about whether a practitioner has conducted an "exposure" is de facto asking whether they completed a series of particular behavioral tasks (e.g., collaboratively created and then reviewed a hierarchy of feared situations, selected

a practice for in-session, identified fear outcome and expectancy, elicited subjective units of distress, coached approach behavior, etc.). The lack of specificity in such a rating item likely contributes to measurement error (i.e., the item is not sufficiently concrete). It therefore impedes a practitioner from self-rating themselves in a behaviorally articulated fashion; it also impedes concordance with any independent observers who have received training on how to characterize each of the particular behavioral tasks under a more articulated coding scheme. The lack of specificity may also make it more difficult for raters to judge quality. Compared to quantity ratings, quality ratings consistently demonstrate lower reliability (e.g., McLeod et al., 2018, McLeod, Martinez, et al., 2022). This is likely because quality ratings are more difficult as they require coders to consider subjective elements (e.g., appropriateness, responsiveness; see Table 1), which can be made more difficult when items lack specificity. Enhancing the specificity of items is likely to facilitate concordance across coders, although this ultimately remains an empirical question to be tested.

Second, the larger discrepancy literature recommends ensuring that the analysis of informant concordance relies on parallel measurement forms. The term "parallel forms" means that measures used by different informants have the same design features (e.g., items, anchors, reporting periods). Many studies included in the review appear to use parallel forms that ask practitioners and independent observers to report on the same items for the same time period (e.g., a treatment session). However, it is not clear in all cases that the informants are given precise instructions about the time period for ratings (e.g., just rate behavior conducted in session). Moreover, not all measures that assess the same content utilize the same rating scales (e.g., a 5- versus a 7-point scale for practitioners and independent observers; McLeod, Sutherland, et al., 2022). Given the importance of ensuring that parallel measurement forms are used, we recommend that future work more carefully report on these design features of the fidelity measures and seek to use parallel measures whenever possible.

In sum, before we can rule out the use of a converging operations conceptualization for fidelity research, more work is needed to determine whether addressing known methodological factors can boost fidelity concordance across informants. In particular, much remains to be done regarding optimally constructing fidelity items to facilitate accurate and consistent assessment across informants. Ideally, such work would proceed in partnership with end users in developing self-report measures (Becker-Haimes et al., 2021); this may dually help enhance the acceptability, appropriateness, and feasibility of developed measures and ultimately improve score validity. Specific strategies include (a) involving stakeholders in all phases of refinement; (b) conducting cognitive interviews with end users to ensure that the item language is appropriate (no jargon, easy to comprehend; Ware et al., 2003); and (c) conducting usability analyses to ensure straightforward interpretation and use of data final-product data, so they can fit into everyday implementation efforts (Mahatody et al., 2010).

Another promising area of research involves reducing measurement error by training informants to be better reporters of fidelity. Compelling examples of this approach can be found in the reviewed studies. For example, Caron and Dozier (2021) provide training and feedback to practitioners asked to code segments of their treatment sessions for protocol fidelity; Hogue et al. (2021) enrolled practitioners in an online coder training course to enhance their self-report acumen regarding EBI delivery with current cases. However, such examples are the exception rather than the rule. Most EBIs fail to require or even describe companion fidelity procedures (Kerns et al., 2021)—this is a first-order procedural failure. And precious few require procedures for training and monitoring practitioners in how to use companion fidelity procedures – this is a second-order, and equally deleterious, procedural failure. It is imperative to train and support practitioners using all fidelity methods they are expected to embrace. To pursue this ambition, we propose that well-established principles for successful EBI training – lean on behavioral rehearsal and active feedback, fortify with ongoing expert consultation, and the like (Frank et al., 2020)—be brought to bear with equivalent rigor in fidelity assessment training.

How Do We Determine if Discrepancies are Noise or Meaningful?

Frameworks (e.g., operations triad) put forth to facilitate the interpretation of multi-informant data in the psychopathology literature (e.g., De Los Reyes et al., 2019) can guide our understanding of multi-informant coders of fidelity and how to determine if discordance is best conceptualized as meaningful information or best attributable to measurement error. In addition to the potential sources of measurement error described above, we posit that there are two understudied, potential sources of variation that could account for fidelity informant discrepancies: (a) biases that each informant brings to the assessment that may yield meaningful information and potential correlates of reporting patterns, and (b)

unique information that a particular informant brings to the assessment.

Concerning biases and reporting patterns, furthering understanding of consistent patterns of agreement or disagreement between fidelity informants and factors influencing those patterns could provide important information for the field about how to optimally integrate fidelity data from multiple informants. An essential step in this work is identifying theoretically consistent biases and candidate correlates that influence rating. As detailed in Table 2, several factors may influence ratings, including but not limited to (a) EBI knowledge (e.g., more knowledge leads to an improved ability to accurately rate); (b) whether the session was emotionally evocative for the practitioner (e.g., the client disclosed a trauma history), making it harder to recall elements of the session; or (c) practitioner self-efficacy with the EBI (e.g., the more confident in one's ability to deliver, the more accurately they may report).

Second, does each informant bring unique information to fidelity assessment? In the psychopathology literature, there is evidence that agreement between different informants (e.g., caregivers and teachers) is impacted by their access to other information (e.g., a child's behavior at home may differ from behavior at school). There are certainly examples where this could be true in fidelity assessment as well, given that different informants often utilize various sources of information (see Table 2). For example, a practitioner might indicate that they spent time rapport building, but the coder using a session recording says it was not present because it happened on the walk from the waiting room to the therapy room. However, the extent to which the operations triad model would refer to this kind of a discrepancy as "domain-relevant information" remains an open question. It heavily influences the interpretation of multi-informant fidelity rating.

Taken together, these concepts raise important questions for future research. Does practitioner self-reported fidelity accurately reflect what was done in session? The answer appears to be no, at least not in most cases, with the fidelity evaluation technology currently in use and the error variance resulting from variations in practitioner training, self-efficacy, and so forth. Regardless, it is de facto what the practitioner perceives themselves to have delivered; such information can be used in conjunction with observation (especially during early learning) to determine how self-reflective and aware of their practice the practitioner is, data that can then be used to guide future consultation/supervision/training. Similarly, suppose we are thinking of client raters, who are probably the least "classically trained" judges. In this case, it is unlikely to expect them to report on fidelity accurately, given their unfamiliarity with how the target interventions are operationalized. However, client perceptions may provide clinically relevant insights into their interpretation of intervention delivery that could be shared with practitioners to facilitate their understanding of their client's perceptions, and to identify gaps where there may be obvious misunderstandings to guide treatment plan refinement.

Looking forward, the field needs to engage in research that can distinguish between error and domain relevance. To accomplish this goal, studies will have to incorporate specific design features. De Los Reyes et al. (2022) noted that a critical first step in making this distinction is identifying what domain-relevant criterion measures should be used for validity studies designed to distinguish between error (i.e., low concordance between raters due to error) and domain-relevance. Here we propose that when treatment fidelity is high, we would likely see the following patterns in this set of domain-relevant criterion measures (Fjermestad et al., 2016; McLeod et al., 2013): lower attrition, greater client participation in therapeutic activities (Chu & Kendall, 2004), stronger youth- and caregiver-practitioner alliance, lower symptomatology, and higher functioning (see Fjermestad et al., 2016; McLeod, Southam-Gerow et al., 2013). With this collection of criterion measures in mind, construct validity tests could be created that evaluate linkages between discrepancies in fidelity informants and these domain-relevant outcomes (e.g., do discrepancies between observer- and self-report fidelity measures predict criterion measures). Another way of distinguishing error from domain-relevance is to conduct incremental validity studies that determine (a) what each fidelity informant predicts, over and above one another (in relation to an independent, domain-relevant criterion variable, such as clinical outcomes) or (b) whether integrated fidelity measurement scores (i.e., discrepancy scores) outperform single fidelity informants in predicting criterion variables. These studies would help rule out the possibility that error is the sole culprit of discrepancies observed in the literature. In other words, the field needs to determine whether multi-informant data can help optimize the prediction of domain-relevant outcomes as part of efforts to determine how best to assess treatment fidelity.

Conclusions About the Current State of the Science

In sum, applying an informant discrepancy lens to fidelity measurement represents a path forward for the field. There is a need for more rigorous research to examine patterns of agreement between different informants. This work should be conducted to decrease measurement

error and value the unique contributions of different informants. Likely, perfect agreement between different informants is not attainable, so understanding what different informants can and cannot tell us about fidelity will allow researchers and implementers to make informed choices about the best-fitting informant(s) and methods for meeting their respective goals.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

Preparation of this article was supported in part by a grant from the Institute of Education Sciences [PI McLeod; R305A210168].

ORCID

Bryce D. McLeod http://orcid.org/0000-0002-0996-0492
Nicole Porter http://orcid.org/0000-0001-6149-9501
Aaron Hogue http://orcid.org/0000-0001-8365-9545
Emily M. Becker-Haimes http://orcid.org/0000-0002-9922-8667
Amanda Jensen-Doss http://orcid.org/0000-0003-4995-7463

References

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

Barber, J. P., Mercer, D., Krakauer, I., & Calvo, N. (1996). Development of an adherence/competence rating scale for individual drug counseling. *Drug and Alcohol Dependence*, *43*(3), 125–132. https://doi.org/10.1016/S03768716(96)013051

Beale, S., Liness, S., & Hirsch, C. R. (2020). Trainee self-assessment of cognitive behaviour therapy competence during and after training. *The Cognitive Behaviour Therapist*, *13*, 13. https://doi.org/10.1017/S1754470X19000357

Becker-Haimes, E. M., Klein, M. R., McLeod, B. D., Schoenwald, S. K., Dorsey, S., Hogue, A., Fugo, P. B., Phan, M. L., Hoffacker, C., & Beidas, R. S. (2021). The TPOCS-self-reported therapist intervention fidelity for youth (TPOCS-SeRTIFY): A case study of pragmatic measure development. *Implementation Research and Practice*, *2*, 2633489521992553. https://doi.org/10.1177/2633489521992553

Becker-Haimes, E. M., Marcus, S. C., Klein, M. R., Schoenwald, S. K., Fugo, P. B., McLeod, B. D., Dorsey, S., Williams, N. J., Mandell, D. S., & Beidas, R. S. (2022). A randomized trial to identify accurate measurement methods for adherence to cognitive-behavioral therapy. *Behavior Therapy*, *53* (6), 1191–1204. Advance online publication. https://doi.org/10.1016/j.beth.2022.06.001

Beidas, R. S., Cross, W., & Dorsey, S. (2014). Show me, don't tell me: Behavioral rehearsal as a training and analogue fidelity tool. *Cognitive and Behavioral Practice*, *21*(1), 1–11. https://doi.org/10.1016/j.cbpra.2013.04.002

Beidas, R. S., Maclean, J. C., Fishman, J., Dorsey, S., Schoenwald, S. K., Mandell, D. S., Shea, J A., McLeod, B D., French, M T., Hogue, A., & Adams, D R. (2016). A randomized trial to identify accurate and cost-effective fidelity measurement methods for cognitive-behavioral therapy: Project FACTS study protocol. *BMC Psychiatry*, *16*(1), 1–10. https://doi.org/10.1186/s12888-016-1034-z

Blackburn, I. M., James, I. A., Milne, D. L., Reichelt, F. K., Standart, S., Garland, A., & Reichelt, F. K. (2001). The Revised Cognitive Therapy Scale (CTSR): Psychometric properties. *Behavioural and Cognitive Psychotherapy*, *29* (4), 431–447. https://doi.org/10.1017/S1352465801004040

Blackburn, I. M., Milne, D. L., & James, I. A. (1997). *How are therapeutic competencies evaluated?* [Paper presentation]. Annual Conference of the British Psychological Society (Division of Clinical Psychology). Edinburgh: Herriot–Watt University.

Breitenstein, S. M., Fogg, L., Garvey, C., Hill, C., Resnick, B., & Gross, D. (2010). Measuring implementation fidelity in a community-based parenting intervention. *Nursing Research*, *59*(3), 158. https://doi.org/10.1097/NNR.0b013e3181dbb2e2

Brookman-Frazee, L., Stadnick, N. A., Lind, T., Roesch, S., Terrones, L., Barnett, M. L., Regan, J., Kennedy, C. A., F Garland, A., & Lau, A. S. (2020). Therapist-observer concordance in ratings of EBP strategy delivery: Challenges and targeted directions in pursuing pragmatic measurement in children's mental health services. *Administration and Policy in Mental Health and Mental Health Services Research*, *48*(1), 1–16. https://doi.org/10.1007/s10488-020-01054-x

Brosan, L., Reynolds, S., & Moore, R. G. (2008). Self-evaluation of cognitive therapy performance: Do therapists know how competent they are? *Behavioural and Cognitive Psychotherapy*, *36*(5), 581–587. https://doi.org/10.1017/S1352465808004438

Caron, E. B., & Dozier, M. (2021). Self-coding of fidelity as a potential active ingredient of consultation to improve clinicians' fidelity. *Administration and Policy in Mental Health and Mental Health Services Research*, *49*(2), 237–254. https://doi.org/10.1007/s10488-021-01160-4

Caron, E. B., Muggeo, M. A., Souer, H. R., Pella, J. E., & Ginsburg, G. S. (2019). Concordance between clinician, supervisor and observer ratings of therapeutic competence in CBT and treatment as usual: Does clinician competence or supervisor session observation improve agreement? *Behavioural and Cognitive Psychotherapy*, *48*(3), 350–363. https://doi.org/10.1017/S1352465819000699

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/10.1093/ptr/8.3.307

Chaffin, M., Hecht, D., Aarons, G., Fettes, D., Hurlburt, M., & Ledesma, K. (2016). EBT fidelity trajectories across training cohorts using the interagency Collaborative Team Strategy.

Administration and Policy in Mental Health and Mental Health Services Research, *43*(2), 144–156. https://doi.org/10.1007/s10488-015-0627-z

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

Chapman, J. E., Sheidow, A. J., Henggeler, S. W., Halliday-Boykins, C. A., & Cunningham, P. B. (2008). Developing a measure of therapist adherence to contingency management: An application of the many-facet rasch model. *Journal of Child & Adolescent Substance Abuse*, *17*(3), 47–68. https://doi.org/10.1080/15470650802071655

Chaturvedi, S. R. B. H., & Shweta, R. C. (2015). Evaluation of inter-rater agreement and inter-rater reliability for observational data: An overview of concepts and methods. *Journal of the Indian Academy of Applied Psychology*, *41*(3), 20–27.

Chevron, E. S., & Rounsaville, B. J. (1983). Evaluating the clinical skills of psychotherapists: A comparison of techniques. *Archives of General Psychiatry*, *40*(10), 1129–1132. https://doi.org/10.1001/archpsyc.1983.01790090091014

Chu, B. C., & Kendall, P. C. (2004). Positive of child involvement and treatment outcome within a manual-based cognitive-behavioral treatment for children with anxiety. *Journal of Consulting and Clinical Psychology*, *72*(5), 821–829. https://doi.org/10.1037/0022-006X.72.5.821

Cicchetti, D. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Earlbaum Associates, Inc.

Comer, J. S., & Kendall, P. C. (2004). A symptom-level examination of parent–child agreement in the diagnosis of anxious youths. *Journal of the American Academy of Child & Adolescent Psychiatry*, *43*(7), 878–886. https://doi.org/10.1097/01.chi.0000125092.35109.c5

Couturier, J., Kimber, M., Barwick, M., McVey, G., Findlay, S., Webb, C., Niccols, A., & Lock, J. (2021). Assessing fidelity to family-based treatment: An exploratory examination of expert, therapist, parent, and peer ratings. *Journal of Eating Disorders*, *9*(1), 1–9. https://doi.org/10.1186/s40337-020-00366-5

De Los Reyes, A. (2011). Introduction to the special section. More than measurement error: discovering meaning behind informant discrepancies in clinical assessments of children and adolescents. *Journal of Clinical Child & Adolescent Psychology*, *40*(1), 1–9. https://doi.org/10.1080/15374416.2011.533405

De Los Reyes, A., Cook, C. R., Gresham, F. M., Makol, B. A., & Wang, M. (2019). Informant discrepancies in assessments of psychosocial functioning in school-based services and research: Review and directions for future research. *Journal of School Psychology*, *74*, 74–89. https://doi.org/10.1016/j.jsp.2019.05.005

De Los Reyes, A., Henry, D. B., Tolan, P. H., & Wakschlag, L. S. (2009). Linking informant discrepancies to observed variations in young children's disruptive behavior. *Journal of Abnormal Child Psychology*, *37*(5), 637–652. https://doi.org/10.1007/s10802-009-9307-3

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin*, *131*(4), 483–509. https://doi.org/10.1037/0033-2909.131.4.483

De Los Reyes, A., Thomas, S. A., Goodman, K. L., & Kundey, S. M. A. (2013). Principles underlying the use of multiple informants' reports. *Annual Review of Clinical Psychology*, *9*(1), 123–149. https://doi.org/10.1146/annurev-clinpsy-050212-185617

De Los Reyes, A., Wang, M., Lerner, M. D., Makol, B. A., Fitzpatrick, W., & Weisz, J. R. (2022). The operations triad model and youth mental health assessments: Catalyzing an paradigm shift in measurement validation. *Journal of Clinical Child & Adolescent Psychology*, 1–36. Advance online publication. https://doi.org/10.1080/15374416.2022.2111684

Dennhag, I., Gibbons, M. B. C., Barber, J. P., Gallop, R., & Crits-Christoph, P. (2012). Do supervisors and independent judges agree on evaluations of therapist adherence and competence in the treatment of cocaine dependence? *Psychotherapy Research*, *22*(6), 720–730. https://doi.org/10.1080/10503307.2012.716528

Dickson, K. S., & Suhrheinrich, J. (2021). Concordance between community supervisor and provider ratings of fidelity: Examination of multi-level predictors and outcomes. *Journal of Child and Family Studies*, *30*(2), 542–555. https://doi.org/10.1007/s10826-020-01877-0

Dorsey, S., Lyon, A. R., Pullmann, M. D., Jungbluth, N., Berliner, L., & Beidas, R. (2017). Behavioral rehearsal for analogue fidelity: Feasibility in a state-funded children's mental health initiative. *Administration and Policy in Mental Health and Mental Health Services Research*, *44*(3), 395–404. https://doi.org/10.1007/s10488-016-0727-4

Fjermestad, K. W., McLeod, B. D., Tully, C. B., & Liber, J. M. (2016). Therapist characteristics and interventions: enhancing alliance and involvement with youth. In S. Maltzman (Ed.), *Oxford handbook of treatment processes and outcomes in counseling psychology* (pp. 97–116). Oxford University Press.

Frank, H. E., Becker-haimes, E. M., & Kendall, P. C. (2020). Therapist training in evidence-based interventions for mental health: A systematic review of training approaches and outcomes. *Clinical Psychology: Science and Practice*, *27*(3), e12330. https://doi.org/10.1111/cpsp.12330

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *9*(3), 330–338. https://doi.org/10.1016/j.sapharm.2012.04.004

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science*, *5*(1), 13–34. https://doi.org/10.1207/S15327841MPEE0501_2

Gresham, F. M., Dart, E. H., & Collins, T. A. (2017). Generalizability of multiple measures of treatment integrity: Comparisons among direct observation, permanent products, and self-report. *School Psychology Review*, *46*(1), 108–121. https://doi.org/10.1080/02796015.2017.12087606

Gumport, N. B., Stephanie, H. Y., Mullin, A. C., Mirzadegan, I. A., & Harvey, A. G. (2019). The validation

of a provider-reported fidelity measure for the transdiagnostic sleep and circadian intervention in a community mental health setting. *Behavior Therapy*, *51*(5), 800–813. https://doi.org/10.1016/j.beth.2019.11.006

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23. https://doi.org/10.20982/tqmp.08.1.p023

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis*, *10*(1), 103–116. https://doi.org/10.1901/jaba.1977.10-103

Henggeler, S. W., Schoenwald, S. K., Liao, J. G., Letourneau, E. J., & Edwards, D. L. (2002). Transporting efficacious treatments to field settings: The link between supervisory practices and therapist fidelity in MST programs. *Journal of Clinical Child & Adolescent Psychology*, *31*(2), 155–167. https://doi.org/10.1207/S15374424JCCP3102_02

Herschell, A. D., Quetsch, L. B., & Kolko, D. J. (2020). 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*, *28*(2), 92–103. https://doi.org/10.1177/1063426618821901

Hill, C. E. (1991). Almost everything you ever wanted to know about how to do process research on counseling and psychotherapy but didn't know who to ask. In C. E. Hill & L. J. Schneider (Eds.), *Research in counseling* (pp. 85–118). Lawrence Erlbaum Associates.

Hogue, A. (2022). Behavioral intervention fidelity in routine practice: Pragmatism Moves to Head of the Class. *School Mental Health*, *14*(1), 1–7. https://doi.org/10.1007/s12310-021-09488-w

Hogue, A., Bobek, M., Dauber, S., Henderson, C. E., McLeod, B. D., & Southam-Gerow, M. A. (2019). Core elements of family therapy for adolescent behavior problems: Empirical distillation of three manualized treatments. *Journal of Clinical Child & Adolescent Psychology*, *48*(1), 29–41. https://doi.org/10.1080/15374416.2018.1555762

Hogue, A., Bobek, M., Porter, N., Dauber, S., Southam-Gerow, M. A., McLeod, B. D., & Henderson, C. E. (2021). Core elements of family therapy for adolescent behavioral health problems: Validity generalization in community settings. *Journal of Clinical Child & Adolescent Psychology*, 1–13. https://doi.org/10.1080/15374416.2021.1969939

Hogue, A., Bobek, M., Porter, N., MacLean, A., Bruynesteyn, L., Jensen Doss, A., Dauber, S., & Henderson, C. E. (2022). Therapist self-report of fidelity core elements of family therapy for adolescent behavior problems: Psychometrics of a pragmatic quality indicator tool. *Administration and Policy in Mental Health and Mental Health Services Research*, *49*(2), 298–311. https://doi.org/10.1007/s10488-021-01164-0

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

Hogue, A., Henderson, C. E., Dauber, S., Barajas, P. C., Fried, A., & Liddle, H. A. (2008). Treatment adherence, competence, and outcome in individual and family therapy for adolescent behavior problems. *Journal of Consulting and Clinical Psychology*, *76*(4), 544–555. https://doi.org/10.1037/0022-006X.76.4.544

Hogue, A., Liddle, H. A., & Rowe, C. (1996). Treatment adherence process research in family therapy: A rationale and some practical guidelines. *Psychotherapy: Theory, Research, Practice, & Training*, *33*(2), 332–345. https://doi.org/10.1037/0033-3204.33.2.332

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

Jensen Doss, A., Cusack, K. J., & De Arellano, M. A. (2008). Workshop-based training in trauma-focused CBT: An in-depth analysis of impact on provider practices. *Community Mental Health Journal*, *44*(4), 227–244. https://doi.org/10.1007/s10597-007-9121-8

Jensen Doss, A., Douglas, S., Phillips, O., Gencdur, D. A., Zalman, A., & Gomez, N. E. (2020). Measurement-based care as a practice improvement tool: Clinical and organizational application in youth mental health. *Evidence-Based Practice in Child and Adolescent Mental Health*, *5*(3), 233–250. https://doi.org/10.1080/23794925.2020.1784062

Kerns, S. E., Perrine, C. M., Sedlar, G., Peterson, R., & Monroe DeVita, M. (2021). Keeping the faith while keeping it real: Practical, empirical approaches to evaluating treatment fidelity. *Global Implementation Research and Applications*, *1*(2), 1–12. https://doi.org/10.1007/s43477-021-00012-5

Koddebusch, C., & Herrmann, C. (2019). Multi-informant assessment of therapeutic competence: Development and initial validation of a set of measurements. *International Journal of Psychology and Psychological Therapy*, *19*(1), 15–28.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*, 155–163.

Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *The American Journal of Psychiatry*, *160*(9), 1566–1577. https://doi.org/10.1176/appi.ajp.160.9.1566

Laird, R. D. (2020). Analytical challenges of testing hypotheses of agreement and discrepancy: Comment on Campione-Barr, Lindell, and Giron (2020). *Developmental Psychology*, *56*(5), 970–977.https://doi.org/10.1037/dev0000763

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement.

Organizational Research Methods, *11*(4), 815–852. https://doi.org/10.1177/1094428106296642

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*, 9. https://doi.org/10.1017/S1754470X16000027

Mahatody, T., Sagar, M., & Kolski, C. (2010). State of the art on the cognitive walkthrough method, its variants and evolutions. *International Journal of Human-Computer Interaction*, *26*(8), 741–85.

Martino, S., Ball, S. A., Nich, C., Frankforter, T. F., & Carroll, K. M. (2008). Community program therapist adherence and competence in motivational enhancement therapy. *Drug and Alcohol Dependence*, *96*(1–2), 37–48. https://doi.org/10.1016/j.drugalcdep.2008.01.020

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

Masia Warner, C., Brice, C., Esseling, P. G., Stewart, C. E., Mufson, L., & Herzig, K. (2013). Consultants' perceptions of school counselors' ability to implement an empirically-based intervention for adolescent social anxiety disorder. *Administration and Policy in Mental Health and Mental Health Services Research*, *40*(6), 541–554. https://doi.org/10.1007/s10488-013-0498-0

Mathieson, F. M., Barnfield, T., & Beaumont, G. (2009). Are we as good as we think we are? Self-assessment versus other forms of assessment of competence in psychotherapy. *The Cognitive Behaviour Therapist*, *2*(1), 43–50. https://doi.org/10.1017/S1754470X08000081

McGrew, J. H., Stull, L. G., Rollins, A. L., Salyers, M. P., & Hicks, L. J. (2011). A comparison of phone-based and on-site assessment of fidelity for assertive community treatment in Indiana. *Psychiatric Services*, *62*(6), 670–674. https://doi.org/10.1176/ps.62.6.pss6206_0670

McGrew, J. H., White, L. M., Stull, L. G., & Wright-Berryman, J. (2013). A comparison of self-reported and phone-administered methods of ACT fidelity assessment: A pilot study in Indiana. *Psychiatric Services*, *64*(3), 272–276. https://doi.org/10.1176/appi.ps.001252012

McLeod, B. D., Cox, J. R., Jensen-doss, A., Herschell, A., Ehrenreich-may, J., & Wood, J. J. (2018). Proposing a mechanistic model of clinician training and consultation. *Clinical Psychology: Science and Practice*, *25*(3), e12260. https://doi.org/10.1111/cpsp.12260

McLeod, B. D., Islam, N. Y., & Wheat, E. (2013). Designing, conducting, and evaluating therapy process research. In J. Comer & P. Kendall (Eds.), *The Oxford handbook of research strategies for clinical psychology* (pp. 142–164). Oxford University Press.

McLeod, B. D., Martinez, R., Southam-Gerow, M. A., Weisz, J. R., & Chorpita, B. F. (2022). Can a single measure estimate protocol adherence for two psychosocial treatments for youth anxiety delivered in community mental health settings? *Behavior Therapy*, *53*(1), 119–136. https://doi.org/10.1016/j.beth.2021.06.008

McLeod, B. D., Smith, M. M., Southam-gerow, M. A., Weisz, J. R., & Kendall, P. C. (2015). Measuring treatment differentiation for implementation research: The Therapy Process Observational Coding System for Child Psychotherapy Revised Strategies Scale. *Psychological Assessment*, *27*(1), 314–325. https://doi.org/10.1037/pas0000037

McLeod, B. D., Southam-gerow, M. A., Tully, C. B., Rodríguez, A., & Smith, M. M. (2013). Making a case for treatment fidelity 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

McLeod, B. D., Southam-Gerow, M. A., & Weisz, J. R. (2009). Conceptual and methodological issues in treatment integrity measurement. *School Psychology Review*, *38*(4), 541–546.

McLeod, B. D., Sutherland, K. S., Broda, M., Granger, K. L., Cecilione, J., Cook, C. R., Conroy, M. A., Snyder, P. A., & Southam-Gerow, M. A. (2022). Examining the correspondence between teacher-and observer-report treatment integrity measures. *School Mental Health*, *14*(1), 20–34. https://doi.org/10.1007/s12310-021-09437-7

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

Moyers, T. B., Martin, T., Manuel, J. K., Miller, W. R., & Ernst, D. (2010). *Revised global scales: Motivational interviewing treatment integrity 3.1. 1 (MITI 3.1. 1)* [Unpublished manuscript]. University of New Mexico.

Mullin, D. J., Saver, B., Savageau, J. A., Forsberg, L., & Forsberg, L. (2016). Evaluation of online and in-person motivational interviewing training for healthcare providers. *Families, Systems, & Health*, *34*(4), 357. https://doi.org/10.1037/fsh0000214

Park, A. L., Chorpita, B. F., Regan, J., Weisz, J. R., & The Research Network on Youth Mental Health. (2015). Integrity of evidence-based practice: Are providers modifying practice content or practice sequencing? *Administration and Policy in Mental Health and Mental Health Services Research*, *42*(2), 186–196. https://doi.org/10.1007/s10488-014-0559-z

Peavy, K. M., Guydish, J., Manuel, J. K., Campbell, B. K., Lisha, N., Le, T., Delucchi, K., & Garrett, S. (2014). Treatment adherence and competency ratings among therapists, supervisors, study-related raters and external raters in a clinical trial of a 12-step facilitation for stimulant users. *Journal of Substance Abuse Treatment*, *47*(3), 222–228. https://doi.org/10.1016/j.jsat.2014.05.008

Perepletchikova, F. (2011). On the topic of treatment fidelity. *Clinical Psychology: Science and Practice*, *18*(2), 148–153. https://doi.org/10.1111/j.1468-2850.2011.01246.x

Perepletchikova, F., & Kazdin, A. E. (2004). Assessment of parenting practices related to conduct problems: Development and validation of the management of children. *Journal of Child and Family Studies*, *13*(4), 385–403. https://doi.org/10.1023/B:JCFS.0000044723.45902.70

Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice*, *12*(4), 365–383. https://doi.org/10.1093/clipsy.bpi045

Regan, J., Park, A., & Chorpita, B. (2019). Choices in treatment integrity: Considering the protocol and consultant recommendations in child and adolescent therapy. *Journal of Clinical Child & Adolescent Psychology*, *48*(sup1), S79–S89.

Rollins, A. L., McGrew, J. H., Kukla, M., McGuire, A. B., Flanagan, M. E., Hunt, M. G., Leslie, D. L., Collins, L. A., Wright-Berryman, J. L., Hicks, L. J., & Salyers, M. P. (2016). Comparison of assertive community treatment fidelity assessment methods: Reliability and validity. *Administration and Policy in Mental Health and Mental Health Services Research*, *43*(2), 157–167. https://doi.org/10.1007/s10488-015-0641-1

Rozek, D. C., Serrano, J. L., Marriott, B. R., Scott, K. S., Hickman, L. B., Brothers, B. M., Lewis, C. C., & Simons, A. D. (2018). Cognitive behavioural therapy competency: Pilot data from a comparison of multiple perspectives. *Behavioural and Cognitive Psychotherapy*, *46*(2), 244–250. https://doi.org/10.1017/S1352465817000662

Sanetti, L. M., Charbonneau, S., Knight, A., Cochrane, W. S., Kulcyk, M. C. M., & Kraus, K. E. (2020). Treatment fidelity reporting in intervention outcome studies in the school psychology literature from 2009 to 2016. *Psychology in the Schools*, *57*(6), 901–922. https://doi.org/10.1002/pits.22364

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

Sheshko, D. M., Lee, C. M., & Gagné, M. H. (2020). Multimethod adherence measurement in an evidence-based parenting program. *Practice Innovations*, *5*(4), 290. https://doi.org/10.1037/pri0000110

Stirman, S. W., Gutner, C. A., Gamarra, J., Suvak, M. K., Vogt, D., Johnson, C., & STRONG STAR Consortium. (2021). A novel approach to the assessment of fidelity to a cognitive behavioral therapy for PTSD using clinical Worksheets: A proof of concept with cognitive processing therapy. *Behavior Therapy*, *52*(3), 656–672. https://doi.org/10.1016/j.beth.2020.08.005

Sutherland, K. S., & McLeod, B. D. (2015a). *Scoring manual for the treatment integrity measure for early childhood settings: the adherence and competence scale* [Unpublished scoring manual prepared]. Virginia Commonwealth University.

Sutherland, K. S., & McLeod, B. D. (2015b). *Scoring manual for the treatment integrity measure for early childhood settings: the school Mental Health teacher report scale* [Unpublished scoring manual prepared]. Virginia Commonwealth University.

Sutherland, K. S., & McLeod, B. D. (2022). Advancing the science of integrity measurement in school mental health research. *School Mental Health*, *14*(1), 1–6. https://doi.org/10.1007/s12310-021-09468-0

Ward, A. M., Regan, J., Chorpita, B. F., Starace, N., Rodriguez, A., Okamura, K., & Research Network on Youth Mental Health. (2013). Tracking evidence-based practice with youth: Validity of the MATCH and standard manual consultation records. *Journal of Clinical Child & Adolescent Psychology*, *42*(1), 44–55. https://doi.org/10.1080/15374416.2012.700505

Ware, N. C., Dickey, B., Tugenberg, T., & McHorney, C. A. (2003). CONNECT: A measure of continuity of care in mental health services. *Administration and Policy in Mental Health and Mental Health Services Research*, *5*(4), 209–221. https://doi.org/10.1023/A:1026276918081