

# Pragmatic fidelity measurement in youth service settings

Implementation Research and Practice  
Volume 4: Jan-Dec 2023 1–13  
© The Author(s) 2023  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/26334895231185380  
journals.sagepub.com/home/irp



Lu Wang<sup>1,2</sup> , Samantha J. Stoll<sup>1</sup>, Christopher J. Eddy<sup>1</sup> ,  
Sarah Hurley<sup>3</sup>, Jocelyn Sisson<sup>3</sup>, Nicholas Thompson<sup>3</sup>,  
Jacquelyn N. Raftery-Helmer<sup>4</sup>, J. Stuart Ablon<sup>1,2</sup> and Alisha R. Pollastri<sup>1,2</sup> 

## Abstract

### Background

Fidelity measurement is critical for developing, evaluating, and implementing evidence-based treatments (EBTs). However, traditional fidelity measurement tools are often not feasible for community-based settings. We developed a short fidelity rating form for the Collaborative Problem Solving (CPS) approach from an existing manualized coding system that requires extensive training. We examined the reliability and accuracy of this short form when completed by trained observers, untrained observers, and self-reporting providers to evaluate multiple options for reducing barriers to fidelity measurement in community-based settings.

### Methods

Community-based treatment providers submitted recordings of youth service sessions in which they did, or did not, use CPS. For 60 recordings, we compared short-form fidelity ratings assigned by trained observers and untrained observers to those provided by trained observers on the manualized coding system. For 141 recordings, we compared providers' self-reported fidelity on the short form to ratings provided by trained observers on the manualized coding system and examined providers' accuracy as a function of their global fidelity.

### Results & Conclusions:

The short form was reliable and accurate for trained observers. An assigned global integrity score and a calculated average of component scores on the short form, but not component scores themselves, were reliable and accurate for observers who had CPS expertise but no specific training on rating CPS fidelity. When providers self-reported fidelity on the short form, their global integrity score was a reliable estimate of their CPS integrity; however, providers with better CPS fidelity were most accurate in their self-reports. We discuss the costs and benefits of these more pragmatic fidelity measurement options in community-based settings.

**Plain Language Summary:** Developing brief, easy-to-use, and reliable tools to measure how well providers deliver evidence-based treatments (EBTs) in community clinical settings is critical to ensure the benefits of EBTs. However, reliable tools are often too time-consuming and not feasible to use in community settings because they require independent observers to receive intensive training on a coding system and to observe live or recorded treatment sessions for reliable and accurate evaluation. This paper describes steps we took to develop a more practical measure of how well providers deliver one EBT, Collaborative Problem Solving (CPS), based on a previously validated measure, to explore whether the

<sup>1</sup>Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA

<sup>2</sup>Department of Psychiatry, Harvard Medical School, Boston, MA, USA

<sup>3</sup>Research Department, Youth Villages, Memphis, TN, USA

<sup>4</sup>Department of Psychology, Worcester State University, Worcester, MA, USA

### Corresponding author:

Alisha Pollastri, Think:Kids in the Department of Psychiatry at Massachusetts General Hospital, 151 Merrimac Street, Boston MA 02114, USA.  
Email: apollastri@mgh.harvard.edu



quality of the measure can be maintained while reducing the need for training independent observers and the need for recording treatment sessions. This work contributes to the growing efforts of developing more pragmatic fidelity measures and introduces a new tool, the CPS Practice Integrity Form (CPS-PIF), as a promising measure for community-based clinical settings using CPS.

### Keywords

treatment fidelity, implementation, mental health, community based, intervention, treatment adherence

## Introduction

Addressing the current youth mental health crisis in the United States (American Academy of Pediatrics, 2021) will depend in part on developing novel interventions, studying their effectiveness, and implementing effective approaches into systems to promote wide dissemination. High-quality methods to measure fidelity are increasingly considered critical for these pursuits (Bond & Drake, 2020). Fidelity, sometimes called integrity, is the extent to which an evidence-based treatment (EBT) is delivered as intended by the developers, and it encompasses several dimensions including (a) adherence, or the extent to which core components are delivered; (b) competence, or the skillfulness with which components are delivered; and (c) differentiation, or the extent to which unique features of the EBT are delivered, making it distinguishable from comparable interventions (Durlak & DuPre, 2008; Perepletchikova et al., 2007; Schoenwald & Garland, 2013). Sound fidelity measures facilitate the identification of core intervention components, allow for the evaluation of whether core components are responsible for treatment effects, structure the process of program implementation, accelerate the uptake of a system-wide practice, enable comparison across sites, and aid scientific communication about the intervention (Bond & Drake, 2020).

Conventional best practices for measuring treatment fidelity include manualized observational coding systems, in which independent observers (who may or may not be experts in the EBT itself) are trained to evaluate the fidelity of live or recorded treatment sessions using a complex rating system. While widely accepted, this approach is time-consuming and not feasible to implement in resource-limited settings (e.g., Hogue, 2022; Perepletchikova, 2011; Schoenwald et al., 2011). Furthermore, practitioners and clients report discomfort with observations and recording, which could result in biased or “artificial” data used for fidelity monitoring (Kimber et al., 2019). As a result, recent studies have sought to identify more *pragmatic* methods for fidelity measurement, i.e., reliable but feasible methods for community-based settings (e.g., Hogue, 2022).

A self-report approach, which does not require recording sessions or additional staff time for observation, would be particularly efficient and likely more acceptable to practitioners and clients. However, previous research suggests

that providers’ self-reports of fidelity often lack concordance with independent observers who are trained to use a manualized coding system (e.g., Caron et al., 2020; Herschell et al., 2020; Hurlburt et al., 2010; McLeod et al., 2022; for exceptions, see Hogue, Bobek, et al., 2022; Hogue et al., 2015). Many factors may contribute to this lack of concordance, including providers’ attribution biases, inaccurate recall, or lack of insight or fidelity to the EBP, which can lead to poor identification of their fidelity errors (Brosan et al., 2008; Caron et al., 2020). Evidence shows that greater fidelity to an EBP is associated with better reliability and accuracy of providers’ self-report. For example, Caron et al. (2020) found smaller discrepancies between self-reported and observers’ fidelity ratings when providers were more competent in the EBT, suggesting that providers’ EBT fidelity may be the key driver of successful self-rating. Thus, any study of providers’ self-reported fidelity should consider the impact of providers’ fidelity on their ability to provide accurate and reliable ratings.

This study sought to develop a more efficient fidelity tool for an EBT that targets youth behavior disorders and, in doing so, extend prior research on pragmatic fidelity measurement. We examined two methods that are practical for community settings: First, we explored whether the new tool yielded reliable and accurate ratings when used by observers who had EBT expertise but had not undergone intensive training on the observational coding system, such as what would be expected of supervisors in community-based clinical settings. We hypothesized that EBT fidelity, rather than familiarity with the coding system, would be critical for the reliability and accuracy of an individual’s fidelity ratings and that these individuals would use the fidelity tool reliably and accurately. Second, we explored whether the tool yielded reliable and accurate ratings when used by providers who self-reported their fidelity following a treatment session and examined the impact of providers’ fidelity on the accuracy of their self-reports. We hypothesized that the accuracy of providers’ self-reports would be positively associated with their fidelity.

By examining the reliability and accuracy of fidelity ratings made by independent observers who have not received training in the manualized coding system and by providers themselves, the current study extends the literature on the utility of pragmatic, lower-cost alternatives for fidelity measurement. Before describing the tool

under investigation, we briefly outline the relevant aspects of the intervention.

## Collaborative Problem Solving

CPS is an evidence-based approach for understanding and treating youth exhibiting challenging behaviors such as oppositionality, aggression, or withdrawal (PracticeWise, 2018). Youth who struggle with executive functioning, communication, and social skills exhibit challenging behavior when they face situations requiring these skills (Greene & Ablon, 2005; Schoemaker et al., 2013; Wang et al., 2019; Zadeh et al., 2007); thus, CPS targets the building of these skills to address the root causes of misbehavior. CPS is associated with improvements in youths' skills and behavior in laboratory and community-based service settings (e.g., Pollastri et al., 2013; Pollastri, Wang, Eddy, et al., 2022), and CPS is currently implemented in hundreds of community-based agencies across North America, Australia, and Europe.

CPS includes three phases: (a) assessment—identify common triggers and expectations that lead to challenging behavior and identifying specific neurocognitive skills that will require improvement for the child to respond more adaptively; (b) planning and prioritization—deciding, for each trigger or unmet expectation, how to respond, which could include pursuing the adult expectation, dropping the expectation temporarily, or solving the problem collaboratively with the child; and (c) intervention—engaging with the child in problem-solving conversations to address triggers and unmet expectations in ways that are mutually agreeable and beneficial. During this problem-solving process, the adult models and supports the youth's practice of taking others' perspectives, oral communication, emotion regulation, maintaining attention, impulse control, and predicting outcomes (Greene & Ablon, 2005). The adult also responds adaptively when lagging skills impact behavior, thereby improving relationships and decreasing conflict (Pollastri et al., 2013).

A fidelity measurement tool called the CPS Manualized Expert-Rated Integrity Coding System (CPS-MEtRICS) was previously developed to support the implementation and study of CPS (Pollastri, Wang, Raftery-Helmer, et al., 2022). In a study conducted with an agency providing in-home services for youth with behavioral problems, the CPS-MEtRICS reliably differentiated between providers delivering CPS versus treatment-as-usual (TAU) and assessed the core components of CPS beyond positive client-provider relationships. An exploratory analysis also suggested that CPS-MEtRICS scores were inversely associated with youth outcomes as measured by critical incidents (see Pollastri, Wang, Raftery-Helmer, et al., 2022). However, the resources needed to train and use the CPS-MEtRICS would be impractical for most community-based settings.

## The Current Study

This study sought to develop a pragmatic fidelity-monitoring tool and methods for its use in community-based settings. The first aim was to design a short form based on the previously validated CPS-MEtRICS that would yield reliable and valid ratings when used by trained CPS-MEtRICS coders (referred to hereafter as "MEtRICS-trained observers"). To accomplish this, we developed the CPS Practice Integrity Form (CPS-PIF) and asked MEtRICS-trained observers to rate fidelity using only the CPS-PIF. After establishing the psychometric properties of the CPS-PIF when used by MEtRICS-trained observers, the second aim was to explore whether the CPS-PIF yielded reliable and accurate ratings when used by observers who had expertise in CPS but who had not been trained as CPS-MEtRICS coders (referred to as "MEtRICS-untrained observers"). The third aim was to explore whether the CPS-PIF yielded reliable and accurate ratings when used by providers as a self-report tool based on their immediate recollection of a session. Comparing results across aims allows us to compare the costs and benefits of each fidelity measurement approach.

## Methods

### Fidelity Measurement Tools

**CPS Manualized Expert-Rated Integrity Coding System (CPS-MEtRICS; Pollastri, Wang, Raftery-Helmer, et al., 2022).** The CPS-MEtRICS is a manualized observational system to measure CPS fidelity. It includes a rating form with 11 items,<sup>1</sup> covering each CPS core component and one global integrity item for the observers' overall impression, as well as a detailed manual with instructions on how to differentiate items, scoring rubrics, and exemplars corresponding to different scores for each item. An independent observer who has expertise in the practice of CPS and who received extensive training on the CPS-MEtRICS rates each component for each 5-min increment of an audio-recorded session, first determining whether each core component is present during that increment and then rating competent adherence on a scale from 1 = *low* to 4 = *high* for any component that was present. Competent adherence is a construct that jointly assesses adherence to and skillful execution of the core components (see Forgatch et al., 2005; Martin et al., 2021; Smith et al., 2016). After coding all 5-min increments, observers assign an overall dichotomous rating regarding the presence/absence of each component during the session and provide a summary competent adherence score for each component present during the session. Finally, observers assign a Global Integrity score (1 to 4), reflecting their clinical judgment of providers' overall CPS fidelity.

**CPS Practice Integrity Form.** The CPS-PIF is an abbreviated version of the CPS-MEtRICS designed by authors of this study. It requires rating the same 11 CPS core

components and global integrity as the CPS-MEtRICS and the same 1 to 4 scale. However, it is simplified in two main ways. First, the manual is replaced by a brief bulleted list of observable behaviors that operationalize each component (see Table 1 for examples). Second, on the CPS-PIF, rather than providing ratings in 5-min increments, raters provide a single competent adherence rating for each present component and global integrity at the end of the session.

## Participants

The study subjects are two subsets of participants from a primary investigation. Study methods are summarized here, and additional information can be found in the publication of the primary study (Pollastri, Wang, Raftery-Helmer, et al., 2022).

Participants are providers and clients from a multiservice agency that serves youth and families across 12 U.S. states through a continuum of care that includes in-home services and who contributed session recordings for fidelity coding. This agency had begun phased CPS training for all staff. West Tennessee (TN) staff had received at least 16 h of didactic training and weekly CPS supervision, while Middle and East TN staff had yet to receive CPS training and were providing TAU.

All in-home service providers from the relevant TN offices were invited to participate. Interested providers

consented to share audio-recorded sessions and employment information, and West TN providers also submitted a self-reported CPS-PIF for each audio-recorded session. Participating providers notified their clients (youth and caregivers) of the opportunity to participate in the study. Interested clients were screened by study staff and consented/assented to sharing audio-recorded sessions and demographic information.

The total sample included 37 providers, with 24 from West TN (CPS) and 13 from Middle and East TN (TAU), and 84 youth receiving in-home services from those providers (59 receiving CPS and 25 receiving TAU). Providers submitted 241 audio-recorded sessions with participating youth (CPS = 159 and TAU = 82); 30% of providers and 50% of youth appeared in two or fewer sessions.

For the first two aims of the current study, in which we evaluated whether the CPS-PIF can yield reliable and accurate fidelity ratings when used by MEtRICS-trained and MEtRICS-untrained observers, we randomly selected 25% of the original 241 sessions from the total sample, including 40 CPS sessions and 20 TAU sessions. Participants in this subsample (Sample A) included 27 providers (24 females and three males; 17 CPS and 10 TAU providers) and 43 youth (ages 4.5 to 17.7, mean age = 12.5; 29 receiving CPS and 14 receiving TAU); 81% of providers and 88% of youth appeared in two or fewer

**Table 1**

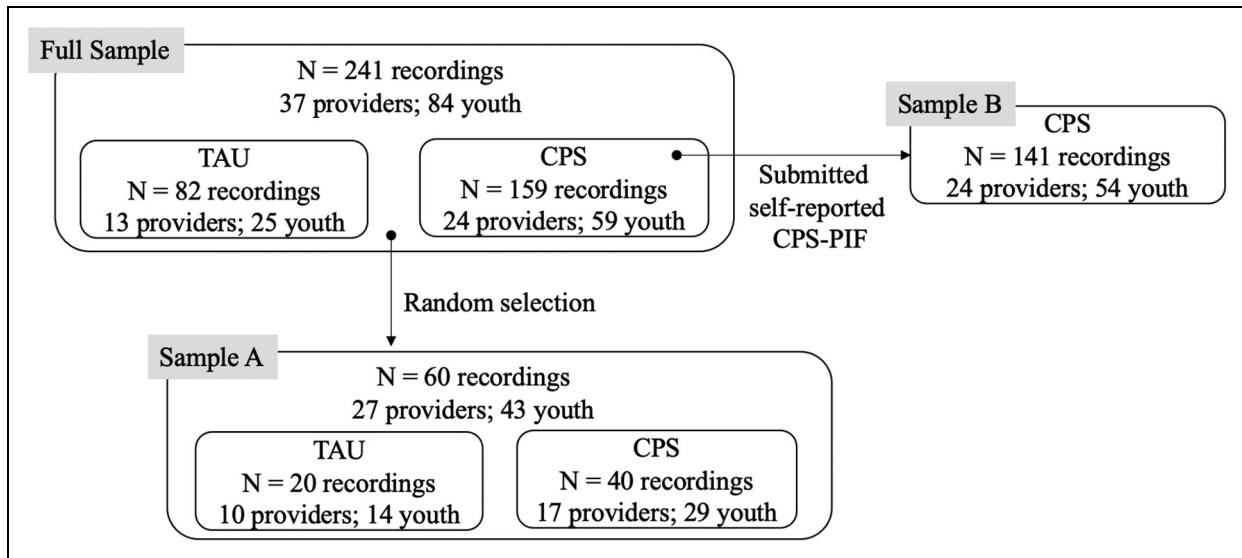
*Sample Content From the CPS-PIF; Three of the 11 CPS Core Components and the List of Observable Behaviors Operationalizing Each Component for the Rater*

Integrity component	Description	Rating
Clarified the youth's concern about a trigger or unmet expectation	<ul style="list-style-type: none"> <li>Used tools: clarifying questions, educated guessing, reflective listening, and reassurance</li> <li>Displayed appropriate empathy and understanding (no drive-by empathy, no over-drilling)</li> <li>Drilled down enough to understand child's concern and learn something new</li> <li>Showed understanding by summarizing before moving on</li> </ul>	1–2–3–4
Shared the adult concern about a trigger or unmet expectation	<ul style="list-style-type: none"> <li>Adult concern was a concern, not a solution</li> <li>Adult concern was about health, safety, learning, or impact on others</li> <li>Adult concern was stated clearly and succinctly</li> <li>Ensured that the youth understood the adult concern without trying to convince the youth that the adult's concern was valid</li> </ul>	1–2–3–4
Brainstormed solutions to address concerns related to the trigger or unmet expectation	<ul style="list-style-type: none"> <li>Summarized the problem as two sets of concerns that need to be resolved</li> <li>Summarized both/all concerns concisely and accurately</li> <li>Encouraged youth to generate solutions first and accepted them for consideration</li> <li>Offered solutions if youth ran out of ideas, but avoided "hijacking" the problem-solving</li> </ul>	1–2–3–4

Note. CPS = Collaborative Problem Solving; PIF = Practice Integrity Form.

**Figure 1**

Relationships Between Samples in the Current Study, Including the Number of Participants in the Treatment-as-Usual (TAU) and Collaborative Problem Solving (CPS) Groups



sessions. No differences in demographics were found between the randomly selected Sample A and the full sample.

For the third aim of the current study, in which we evaluated whether a self-reported CPS-PIF can yield reliable and accurate fidelity ratings, we selected only the sessions recorded by CPS providers who also submitted a self-report CPS-PIF with their session recording ( $N=141$ ). Participants in this subsample (Sample B) included 24 providers (22 females and two males) and 54 youth (ages 6.1 to 17.4, mean age = 12.2); 42% of providers and 61% of youth appeared in two or fewer sessions.

The designation of these samples is illustrated in Figure 1. The demographics of providers and participating youth are summarized in Table 2.

## Procedure

The relevant Institutional Review Board approved all procedures. Participating providers audio-recorded treatment sessions with participating clients and securely submitted recordings to study staff. CPS-trained providers from West TN also submitted a self-reported CPS-PIF after recording a treatment session. Providers and participating families received a small gift card per submitted audio recording.

Each of the 241 sessions in the total sample was coded by two MEtRICS-trained observers using the CPS-MEtRICS as part of the primary investigation (for the complete procedure, see Pollastri, Wang, Raftery-Helmer, et al., 2022). The average MEtRICS score across two MEtRICS-trained observers is used in the current study as the benchmark. MEtRICS scores from the 60 recordings randomly selected

for Sample A did not differ from the overall sample ( $ps > .05$  for both the TAU and the CPS groups). Sample A recordings ranged from 15 to 62 min (mean = 41) and had adequate audio quality to obtain a confident fidelity rating.

To evaluate whether the CPS-PIF could be used reliably and accurately, MEtRICS-trained and MEtRICS-untrained observers completed a CPS-PIF for each of the 60 recordings in Sample A. These were compared to the benchmark. All observers were blind to group (CPS or TAU) for their assigned recordings. CPS-PIF self-report ratings submitted by providers at the time of the recording were also compared to the benchmark.

### CPS-PIF Ratings From MEtRICS-Trained Observers

All seven MEtRICS-trained observers from the primary investigation were invited to provide CPS-PIF ratings in the current study, and five accepted. These five MEtRICS-trained observers had received 36 h of didactic training and at least 12 h of CPS coaching, and all had been practicing and supervising CPS for at least 2 years. In addition, all had previously demonstrated good to excellent inter-rater reliability using the CPS-MEtRICS. For the current study, each trained observer provided CPS-PIF fidelity ratings for 12 session recordings that they had not rated in the prior study.

### CPS-PIF Ratings From MEtRICS-Untrained Observers

Ten individuals employed by the same clinical agency with the same amount of CPS training and clinical experience as MEtRICS-trained observers but who had not been trained on the CPS-MEtRICS were invited to provide CPS-PIF ratings for the current study, and all accepted. Each untrained observer provided fidelity ratings for six

**Table 2**  
Participant Demographics

Providers	Full sample (N = 37)	Sample A (N = 27)	Sample B (N = 24)
Sex	33 females (89%) 4 males (11%)	24 females (89%) 3 males (11%)	22 females (92%) 2 males (8%)
Race	21 White (57%) 13 Black (35%) 3 Other (8%)	16 White (59%) 10 Black (37%) 1 Other (4%)	11 White (46%) 11 Black (46%) 2 Other (8%)
Mean employment length	2.2 years	2.2 years	2.2 years
Highest degree	16 Bachelor's (43%) 21 Master's (57%)	10 Bachelor's (37%) 17 Master's (63%)	9 Bachelor's (38%) 15 Master's (62%)
Youth	Full sample (N = 84)	Sample A (N = 43)	Sample B (N = 54)
Sex	31 females (37%) 53 males (63%)	18 females (42%) 25 males (58%)	17 females (31%) 37 males (69%)
Race	35 White (42%) 34 Black (40%) 15 Other (18%)	22 White (51%) 18 Black (42%) 3 Other (7%)	11 White (20%) 31 Black (57%) 12 Other (23%)
Mean age	12.6 years	12.5 years	12.4 years
Mean length of service	116 days	124 days	117 days
Diagnostic category			
Attention deficit/ hyperactivity disorder	21 (25%)	7 (16%)	17 (31%)
Disruptive behavior disorders	16 (19%)	9 (21%)	8 (15%)
Depressive disorders	16 (19%)	10 (23%)	10 (19%)
Episodic mood disorders	14 (17%)	6 (14%)	10 (19%)
Adjustment disorders	5 (6%)	3 (7%)	5 (9%)
Anxiety disorders	4 (5%)	2 (5%)	1 (2%)
Posttraumatic stress disorders	3 (4%)	1 (2%)	2 (4%)
Autism spectrum disorder	2 (2%)	1 (2%)	0
Psychotic disorder	2 (2%)	2 (5%)	1 (2%)

randomly selected session recordings and was blind to the treatment used.

## Analysis Plan

Comparisons between forms (CPS-MEtRICS and CPS-PIF by trained and untrained observers and self-report) include component scores and two types of summary scores: a global integrity score assigned by the observer, and an average integrity score, calculated as the mean competent adherence for all present components in a session. The correlations between these two summary scores range from 0.70 to 0.86 across different forms.

Reliability, or the extent to which observer- or self-ratings on the CPS-PIF covary with previously validated ratings on the CPS-MEtRICS, is operationalized by inter-rater reliability coefficients (ICCs). In the current study, to measure the reliability of the observer-rated CPS-PIF (Aims 1 and 2), ICC(2,1) was calculated using a two-way random effect model based on a single rating and absolute agreement (Koo & Li, 2016). To measure the reliability of the self-reported CPS-PIF (Aim 3), ICC(1,1) was calculated using a one-way random effect model based on a single rating and absolute agreement (see Brookman-Frazer et al., 2021). We

used Cicchetti's (1994) standards, whereby ICCs below 0.40 were poor, ICCs between 0.40 and 0.59 were fair, ICCs between 0.60 and 0.74 were good, and ICCs above 0.74 indicated excellent agreement.

Accuracy, or the extent of the match between observer- or self-rated competent adherence on the CPS-PIF and previously validated ratings on the CPS-MEtRICS, was operationalized by comparing the competent adherence ratings for each component, global integrity and average integrity across forms, and by exploring whether ratings on the CPS-PIF differentiated CPS and TAU sessions (Aims 1 to 3). Multilevel models (session nested under youth, further nested under providers) with forms and treatment groups as predictors were conducted to examine the differences between global and average integrity scores on the CPS-MEtRICS and the CPS-PIF when MEtRICS-trained and untrained observers rated the CPS-PIF. Multilevel models with forms as the predictor were conducted to examine the differences between global and average integrity scores on the CPS-MEtRICS and the self-reported CPS-PIF. All the multilevel models included random intercepts for level 2 (youth) and level 3 (providers) factors, and no random slopes for form or treatment group were estimated (the model failed to converge due to sample size). Restricted

maximum likelihood method was used for unbiased estimates with a small sample. Given that a component may not occur in each session (the median occurrence of the 11 components ranged from 45% to 57% across forms and is the lowest in TAU sessions), paired *t*-tests were used to examine the differences in each component. Following prior practice (Hogue, Bobek, et al., 2022), we adjusted the alpha level to  $p < .01$  to reduce the likelihood of Type I error in these component-wise analyses.

Additionally, to explore whether the accuracy of providers' self-reports varied by providers' fidelity, all CPS providers were divided into high- and low-fidelity groups based on the average global integrity scores assigned by MEtRICS-trained observers in the primary investigation. Those with average global integrity scores of 2.5 or higher were categorized as high fidelity ( $n = 12$ ), while those whose average global integrity scores were below 2.5 were classified as low fidelity ( $n = 12$ ), resulting in 69 recordings from low-fidelity providers and 72 recordings from high-fidelity providers. Repeated 2 (form) by 2 (fidelity) ANOVAs on self-reported global and average integrity were used to explore interactions between providers' fidelity and the accuracy of their self-report. Results using continuous fidelity scores were consistent with dichotomized fidelity; the latter was reported for ease of interpretation and clinical implications.

## Results

### Aim 1: CPS-PIF Used by MEtRICS-Trained Observers

As shown in Table 3, the ICCs between MEtRICS-trained observers' ratings on the CPS-MEtRICS and the CPS-PIF were good for 6 out of the 11 components (ICCs = 0.63–0.73), fair for four components (ICCs = 0.44–0.58), poor

for one component, and excellent for global integrity (ICC = 0.75).

As shown in Table 4, MEtRICS-trained observers' mean ratings were similar across forms, especially for CPS sessions in which no significant differences were found. For TAU sessions, MEtRICS-trained observers rated only one item (Item 11), lower on the CPS-PIF,  $t(19) = -3.39$ ,  $p < .01$ . Furthermore, MEtRICS-trained observers assigned higher competent adherence scores for CPS providers compared to TAU providers for three of the nine CPS-PIF components that could be compared between treatment groups (remaining components did not occur frequently enough for comparison). Multilevel models with forms and treatment groups as predictors for global and average integrity revealed no differences between the CPS-MEtRICS and CPS-PIF when rated by MEtRICS-trained observers (marginal  $r^2 = .00$  for both), and there were no interactions between forms and treatment group. For the CPS-PIF alone, CPS providers received higher global integrity scores (marginal  $r^2 = .23$ ) and average scores (marginal  $r^2 = .11$ ) than TAU providers.

### Aim 2: CPS-PIF Used by MEtRICS-Untrained Observers

As shown in Table 3, the ICCs between MEtRICS-trained observers' ratings on the CPS-MEtRICS and MEtRICS-untrained observers' ratings on the CPS-PIF were good for one of the 11 components (ICC = 0.61), fair for six components (ICCs = 0.4–0.56), poor for four components, and excellent for global integrity (ICC = 0.77).

As shown in Table 4, MEtRICS-untrained observers assigned lower scores for several components on the

**Table 3**

*Inter-Rater Agreement (ICC and 95% CI) Between Ratings on the CPS-MEtRICS and Three Types of CPS-PIF Raters: MEtRICS-Trained Observers, MEtRICS-Untrained Observers, and Providers' Self-Report*

Item	CPS-PIF by MEtRICS- trained observers	CPS-PIF by MEtRICS-untrained observers	CPS-PIF by providers' self-report
1	0.63 (0.45–0.76)	0.56 (0.36–0.71)	0.16 (–0.17–0.4)
2	0.48 (0.27–0.65)	0.56 (0.36–0.71)	0.43 (0.21–0.59)
3	0.58 (0.39–0.73)	0.43 (0.15–0.64)	–0.37 (–0.91–0.02)
4	0.53 (0.32–0.69)	0.5 (0.25–0.68)	0.42 (0.19–0.58)
5	0.44 (0.22–0.62)	0.33 (0.01–0.57)	0.48 (0.27–0.63)
6	0.68 (0.52–0.8)	0.61 (0.42–0.75)	0.39 (0.15–0.56)
7	0.65 (0.47–0.77)	0.20 (–0.04–0.42)	0.62 (0.47–0.73)
8	0.73 (0.57–0.84)	0.48 (0.26–0.65)	0.51 (0.32–0.65)
9	0.30 (0.05–0.51)	0.14 (–0.12–0.38)	0.20 (–0.11–0.43)
10	0.64 (0.46–0.77)	0.40 (0.16–0.59)	0.31 (0.04–0.51)
11	0.65 (0.43–0.79)	0.34 (–0.05–0.62)	0.53 (0.35–0.66)
Global	0.75 (0.61–0.84)	0.77 (0.58–0.87)	0.68 (0.56–0.77)

Note. CPS = Collaborative Problem Solving; PIF = Practice Integrity Form; MEtRICS = Manualized Expert-Rated Integrity Coding System.

**Table 4**  
**Comparing MERICS-Trained Observers' Competent Adherence Ratings on the CPS-MERICS to MERICS-Trained and MERICS-Untrained Observers' Ratings on the CPS-PIF, for CPS and TAU Sessions**

Item	1. CPS-MERICS			2. CPS-PIF by trained observers			3. CPS-PIF by untrained observers			1 versus 2		1 versus 3	
	Mean (SD)			Mean (SD)			Mean (SD)			Paired-t or multilevel model		Paired-t or multilevel model	
	TAU	CPS		TAU	CPS		TAU	CPS		TAU	CPS	TAU	CPS
1	1.88 (0.63)	3.06 (0.85)		1.50 (0.71)	3.00 (0.74)		1.00 (-)	2.09 (0.97)		—	-1.71	—	-3.62*
2	—	3.25 (0.66)		3.00 (-)	2.58 (0.67)		1.25 (0.50)	2.36 (1.26)		—	-2.25	—	-1.22
3	—	2.28 (0.83)		—	3.00 (0.82)		1.00 (-)	2.50 (1.08)		—	1.22	—	2.20
4	1.59 (0.70)	2.61 (0.99)		1.60 (0.89)	2.77 (0.92)		1.25 (0.50)	2.30 (1.05)		-3.00	-0.08	1.00	-1.74
5	1.71 (0.50)	2.63 (0.95)		2.00 (0.89)	2.72 (0.89)		1.00 (-)	2.09 (0.99)		0.83	0.36	-1.00	-2.69
6	1.38 (0.75)	2.88 (0.76)		1.80 (0.84)	2.66 (1.08)		1.25 (0.50)	2.37 (1.13)		—	-1.34	—	-1.55
7	1.12 (0.34)	2.07 (1.04)		1.14 (0.38)	2.23 (0.94)		1.50 (0.58)	2.26 (1.04)		-1.00	0.65	2.00	1.70
8	1.38 (0.52)	2.50 (1.05)		1.00 (-)	2.17 (1.03)		1.00 (0)	2.27 (1.12)		-3.00	-1.83	—	-1.98
9	2.00 (0.76)	2.65 (1.18)		2.44 (1.24)	2.29 (1.07)		1.25 (0.50)	2.09 (0.91)		0.00	-0.87	—	-2.00
10	2.95 (1.47)	3.55 (1.09)		2.50 (1.54)	3.55 (1.09)		2.50 (1.54)	3.48 (1.15)		-1.37	0.00	-1.14	-0.37
11	2.35 (0.93)	3.17 (0.76)		1.65 (0.67)	2.98 (0.97)		1.20 (0.62)	2.25 (1.13)		-3.39*	-1.95	-4.83**	-6.02**
Avg	2.00 (0.60)	2.78 (0.68)		1.97 (0.81)	2.71 (0.74)		1.82 (0.84)	2.43 (0.88)		-0.46	—	-3.00*	—
Glob	1.25 (0.38)	2.62 (0.85)		1.45 (0.51)	2.52 (0.88)		1.15 (0.37)	2.21 (1.02)		0.10	—	-3.71**	—

Note. CPS = Collaborative Problem Solving; PIF = Practice Integrity Form; MERICS = Manualized Expert-Rated Integrity Coding System; TAU = treatment-as-usual; Trained or Untrained Obs. = MERICS-trained or MERICS-untrained observers; Avg = average integrity score (calculated); Glob = global integrity score (assigned).

\*  $p < .01$ ; \*\*  $p < .001$ ; †  $p < .05$ .



CPS-PIF than MEtRICS-trained observers' ratings on the CPS-MEtRICS. These differences were significant for one component for TAU providers and two components for CPS providers. MEtRICS-untrained observers assigned higher competent adherence scores for CPS compared to TAU providers for six of the 11 components. For global and average integrity, multilevel models with forms and treatment groups as predictors revealed that MEtRICS-untrained observers' scores on the CPS-PIF were slightly but significantly lower than MEtRICS-trained observers' scores on the CPS-MEtRICS (marginal  $r^2 = .02$  and  $.03$ , respectively), and there were no interactions between forms and the treatment group. When the CPS-PIF was rated by MEtRICS-untrained observers, CPS providers received significantly higher global integrity scores than those for TAU providers (marginal  $r^2 = .24$ ). However, the difference in average integrity scores was not significant (marginal  $r^2 = .05$ ).

### Aim 3: CPS-PIF Used by Providers for Self-Report

As shown in Table 3, the ICCs between MEtRICS-trained observers' ratings on the CPS-MEtRICS and providers' self-reported CPS-PIF were good for one component (ICC = 0.62), fair for five components (ICCs = 0.42–0.53), poor for five components, and good for global integrity (ICC = 0.68).

As shown in Table 5, providers using the self-reported CPS-PIF assigned significantly higher competent adherence scores for five components and significantly lower scores for one component compared to MEtRICS-trained observers' ratings on the CPS-MEtRICS. For global and average integrity, multilevel models with form as the predictor revealed no difference in providers' self-rated global integrity on the CPS-PIF compared to MEtRICS-trained observers' ratings on the CPS-MEtRICS (marginal  $r^2 = .00$ ), but the average integrity score was significantly higher when self-rated on the CPS-PIF (marginal  $r^2 = .02$ ).

Finally, multilevel models with form (CPS-MEtRICS or CPS-PIF self-report) and providers' fidelity (low or high) were used to examine whether the accuracy of providers' self-reported CPS-PIF varied by providers' fidelity as rated by an independent observer. Results suggest a significant interaction for both global integrity,  $t(138) = 2.53$ ,  $p < .05$ , and average integrity,  $t(299) = 3.89$ ,  $p < .001$ . Multilevel post-hoc analyses conducted separately for high-fidelity and low-fidelity providers suggested that low-fidelity providers' self-reported global integrity scores were significantly higher than MEtRICS-trained observers' ratings on the CPS-MEtRICS (mean = 2.36 vs. 2.11);  $t(127) = 2.47$ ,  $p < .05$ . For providers with high fidelity, there were no differences between global integrity on the self-reported CPS-PIF (mean = 3.01) and MEtRICS-trained observers' ratings on the CPS-MEtRICS (mean = 3.10);  $t(124) = -0.74$ ,  $p = \text{n.s.}$  The

**Table 5**

Comparing MEtRICS-Trained Observers' Competent Adherence Ratings on the CPS-MEtRICS to Providers' Self-Reported CPS-PIF

Item	1.	2. CPS-PIF by	1 versus 2 Paired-t or multilevel model
	CPS-MEtRICS	self-report	
	Mean (SD)	Mean (SD)	
1	3.29 (0.75)	2.78 (0.89)	1.56
2	2.93 (0.87)	2.79 (0.95)	1.20
3	2.25 (0.79)	2.60 (0.90)	-1.57
4	2.6 (0.83)	2.82 (0.89)	-2.92*
5	2.47 (0.84)	2.97 (0.81)	-4.99**
6	2.58 (0.86)	2.92 (0.76)	-3.67**
7	1.99 (0.94)	2.80 (0.92)	-7.73**
8	2.25 (0.96)	2.80 (0.93)	-6.58**
9	2.44 (1.11)	2.97 (0.91)	-1.79
10	3.28 (1.29)	3.34 (1.25)	-0.47
11	3.18 (0.81)	2.90 (0.88)	3.53*
Avg.	2.79 (0.71)	2.98 (0.67)	-3.26*
Glob.	2.62 (0.81)	2.70 (0.78)	-1.37

Note. CPS = Collaborative Problem Solving; PIF = Practice Integrity Form; MEtRICS = Manualized Expert-Rated Integrity Coding System; Avg = average integrity score (calculated); Glob = global integrity score (assigned).

\*  $p < .01$ ; \*\*  $p < .001$ .

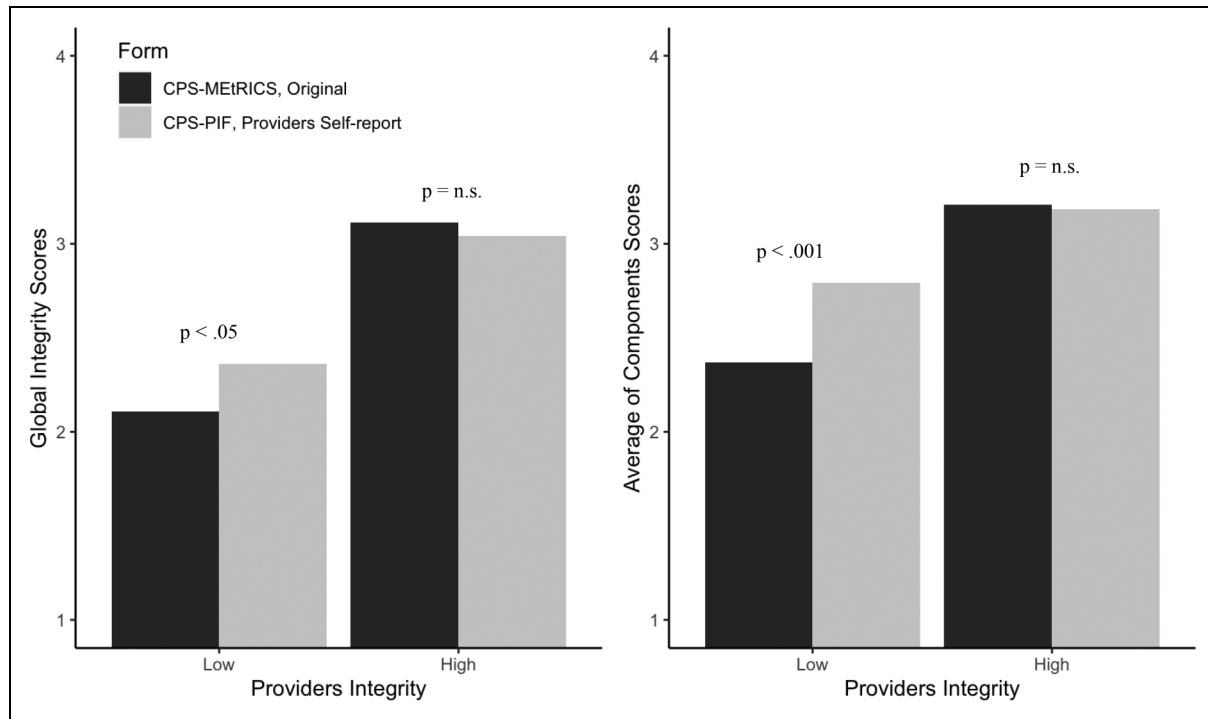
same patterns were found for average integrity, with significant differences between forms for providers with low fidelity (mean = 2.37 vs. 2.79),  $t(111) = 4.51$ ,  $p < .001$ , and not for providers with high fidelity (mean = 3.21 vs. 3.18);  $t(70) = -0.41$ ,  $p = \text{n.s.}$  See Figure 2.

## Discussion

Reliable and valid fidelity tools are considered critical for intervention development, evaluation, and implementation, but traditional measures and methods are not practical for use in community-based settings (Hogue, 2022; Stirman, 2020). To address common barriers, we developed a simplified fidelity rating tool, the CPS-PIF, based on an existing manualized observational coding system for CPS (CPS-MEtRICS) and evaluated three methods for its use. We established the reliability and validity of the CPS-PIF when used by MEtRICS-trained observers. Then we examined whether it could be used reliably and accurately by MEtRICS-untrained observers and by CPS providers as a self-report tool. Our findings suggest that the CPS-PIF can serve as a simpler alternative to the CPS-MEtRICS for MEtRICS-trained observers. The global integrity score and a calculated average of component scores can also be used reliably by observers with CPS expertise but no specific training on rating CPS fidelity. When providers self-report their integrity on the CPS-PIF, the global integrity score can be used to estimate their CPS integrity; notably, this estimate is most accurate for providers independently rated as having good CPS fidelity.

**Figure 2**

Interaction Between Form and Providers' Integrity Using Global Integrity Score and Average Integrity Score.



Note. CPS = Collaborative Problem Solving; MEtRICS = manualized expert-rated integrity rating system; PIF = Practice Integrity Form; n.s. = not significant

We recommend replacing the CPS-MEtRICS with the CPS-PIF when resources are available for training on the coding system and when component-level fidelity is required, based on three facts: MEtRICS-trained observers yielded reliable CPS-PIF ratings for 10 out of 11 components and global CPS integrity, the mean ratings on the CPS-PIF did not differ from the mean ratings on the CPS-MEtRICS, and ratings on the CPS-PIF successfully differentiated between providers of CPS and TAU (Pollastri, Wang, Raftery-Helmer, et al., 2022). Using the CPS-PIF provides a slight increase in efficiency compared to using the CPS-MEtRICS, decreasing the time required to rate each 50-min session from 1.5 h to below 1 h (see Table 6). However, due to the cost of observers' training and observation time, this is likely to be an impractical solution in most community-based settings.

We had hypothesized that observers who were experts in CPS but did not receive extensive training on the coding system would provide accurate ratings on the CPS-PIF, but this was only partially true. The accuracy of fidelity ratings for specific components was inadequate; however, the average of component scores and the assigned global integrity scores were accurate proxies for the CPS-MEtRICS. If a global integrity score is all that is needed to monitor the quality of clinical care, guide

ongoing supervision, or inform program leaders on the allocation of training resources, the CPS-PIF can serve as a pragmatic and cost-saving alternative to the CPS-MEtRICS (see Table 6). For example, for an agency in which two clinical supervisors are assigned as observers and where 50 providers each submit two recorded sessions annually, using the CPS-MEtRICS would require 90 h of training time at startup and 150 h of observer time per year (at an annual salary of \$80,000 and a three-year staffing cost of \$21,600). In contrast, using the CPS-PIF by untrained observers would require no training time at startup and 100 h of observer time per year (a three-year staffing cost of \$12,000 and a cost savings of \$9,600).

CPS providers' self-reports reached adequate reliability for only half the components on the CPS-PIF, but self-reported global integrity was reliable and accurate, consistent with prior studies on other EBTs. Also consistent with prior research in this area (e.g., Caron et al., 2020), providers who were less competent in CPS tended to overestimate their global integrity. Nevertheless, the tradeoff in self-report accuracy may be worth the cost-savings (see Table 6), which would be in line with the growing trend of using providers' self-report for pragmatic fidelity evaluation (e.g., Hogue, Bobek, et al., 2022; Hogue et al., 2015).

**Table 6**  
*Comparison of Staff Time Required for Each of Four Methods of CPS Integrity Measurement*

Task	Attend training on the coding system	Observe the recorded session	Provide ratings for every 5-min of observation	Provide summary ratings at the end of observation
CPS-MEtRICS	45 h per new coder	1 h per session	25 min per session	5 min per session
CPS-PIF coded by MEtRICS-trained observers	45 h per new coder	1 h per session	None	5 min per session
CPS-PIF coded by MEtRICS-untrained observers	None	1 h per session	None	5 min per session
CPS-PIF self-rated by provider after session	None	None	None	5 min per session

Note. CPS = Collaborative Problem Solving; MEtRICS = Manualized Expert Rated Integrity Coding System; PIF = Practice Integrity Form.

Beyond providing new methods for evaluating the fidelity of CPS, this study also serves as a model for fidelity tool development and provides some generalizable information and future directions. First and most importantly, it suggests that training observers in structured fidelity assessment may be necessary if precise information is needed about the integrity of individual treatment components or for research, but when a general sense of global fidelity is all that is needed, there may be several more pragmatic options.

Second, our results suggest that two ingredients contribute to the maximal accuracy of an observer's fidelity ratings: expertise in the treatment approach and specific training in assessing fidelity to the approach. Despite observers with CPS expertise in both our trained and untrained groups, those with training in the integrity coding system exhibited more robust reliability. This is in contrast with the assertion that fidelity-trained observers outperform untrained observers simply because training increases the observer's knowledge of the EBT (see Caron & Dozier, 2022). While our CPS experts were unlikely to gain more knowledge of the approach itself, it appears that training on the coding system, which includes reading and discussing exemplars and receiving feedback on one's ratings of mock sessions, changed the observers' sensitivity to the nuances of CPS practice beyond what CPS expertise alone could provide. This finding can inform the development of activities to increase practitioners' fidelity (Caron & Dozier, 2022; Hogue, Porter, et al., 2022).

Finally, consistent with other studies, self-reported integrity was more accurate when providers had higher fidelity to the approach, as rated by an independent observer. This suggests that clinical leaders can trust that providers with self-rated low fidelity need more support but that they should treat self-reported high integrity with caution, perhaps following up with independent observation. It is currently unknown why providers with lower fidelity overestimate their abilities and whether this problem can be remedied. Past suggestions have included that providers

need sufficient expertise in an approach to recognize their own fidelity mistakes or that self-ratings reflect intentions as well as actual behaviors (Brookman-Frazee et al., 2021). It may also be that providers' memories of sessions are generally biased or inaccurate. Several methods have been explored to improve the reliability and accuracy of providers' self-reports, for example, training providers with exemplars, mock ratings of recorded treatment sessions, and fidelity feedback (e.g., Caron & Dozier, 2022; Hogue, MacLean, et al., 2022; Hogue, Porter, et al., 2022). Other ideas include using structured notetaking during sessions or providing abbreviated recordings to increase the recall accuracy. This will be an important area for continued research on fidelity assessment.

Several limitations and other factors should be considered when interpreting these results. For instance, while collecting these data in a naturalistic community setting was strength of the study, it also resulted in missing data and a more complex nesting structure than is typically found in laboratory research. Additionally, it is still being determined whether we can generalize results across interventions since fidelity requirements for different EBTS can be quite heterogeneous. Strict integrity may be more critical for some interventions than others. Finally, while global and average integrity scores were strongly correlated for untrained observers in this study, whether they can be used interchangeably warrants more study. An average score could be considered more robust, but each component in the average conveys equal weight on the summary score, which may or may not be theoretically consistent.

Despite some limitations and remaining questions, designing pragmatic tools to monitor fidelity is critical as our field seeks to develop effective interventions and translate evidence-based approaches into community-based clinical care. Moreover, a recent meta-analysis suggests that while there is an overall weak relationship between fidelity and outcomes across studies of youth interventions, there is significant heterogeneity across interventions and studies, and further research is needed (Collyer et al.,

2020). Like with CPS, where prior evidence suggests an inverse relationship between fidelity and youths' critical incidents (Pollastri, Wang, Raftery-Helmer, et al., 2022), further studies examining the relationship between EBT fidelity and outcomes will require measurement tools that are both reliable and pragmatic for use in the community-based settings where they are implemented. Future research should explore extending the current findings on fidelity measurement to other EBTs.

### Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article. At the time of publication of this article, the authors Ablon and Pollastri were employed by Think:Kids, a program at Massachusetts General Hospital that contracts with organizations to train and implement the CPS approach.

### Acknowledgments


All authors would like to express their sincere appreciation to the coders who generated much of the data analyzed in this article: Brent Doyle, Bonnie McKinney, Kathleen McNamara, Sara Skonieczny, Rebecca Smith, Rhonda Stempkowski, Lauren Wantz, and Katherine Peatross, as well as the providers and families who allowed us to record their sessions to answer these important questions. Without them, this project would not have been possible.


### Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

### ORCID iDs

Lu Wang  <https://orcid.org/0000-0001-8595-8194>

Christopher J. Eddy  <https://orcid.org/0000-0003-0027-5496>

Alisha R. Pollastri  <https://orcid.org/0000-0001-9368-7128>

### Note

1. The original CPS-METRICS included 14 items on CPS core components and one global integrity item. Three items on CPS core components had poor inter-rater reliability when used by METRICS-trained observers using the CPS-METRICS in a previous study. These items were kept in the form for further monitoring but were dropped in all analyses.

### References

- American Academy of Pediatrics. (2021). *AAP-AACAP-CHA Declaration of a National Emergency in Child and Adolescent Mental Health*. <https://www.aap.org/en/advocacy/child-and-adolescent-healthy-mental-development/aap-aacap-cha-declaration-of-a-national-emergency-in-child-and-adolescent-mental-health/>.
- Bond, G. R., & Drake, R. E. (2020). Assessing the fidelity of evidence-based practices: History and current status of a standardized measurement methodology. *Administration and Policy in Mental Health and Mental Health Services Research*, 47(6), 874–884. <https://doi.org/10.15585/mmwr.su7102a1>
- Brookman-Frazee, L., Stadnick, N. A., Lind, T., Roesch, S., Terrones, L., Barnett, M., Regan, J., Kennedy, C. A., Garland, A. F., & Lau, A. S. (2021). Therapist-observer concordance in ratings of EBP strategy delivery: Challenges and targeted directions in pursuing pragmatic measurement in children's mental health services. *Administration and Policy in Mental Health*, 48(1), 155–170. <https://doi.org/10.1007/s10488-020-01054-x>
- Brosan, L., Reynolds, S., & Moore, R. G. (2008). Self-evaluation of cognitive therapy performance: Do therapists know how competent they are? *Behavioural and Cognitive Psychotherapy*, 36(5), 581–587. <https://doi.org/10.1017/S1352465808004438>
- Caron, E. B., & Dozier, M. (2022). Self-coding of fidelity as a potential active ingredient of consultation to improve clinicians' fidelity. *Administration and Policy in Mental Health and Mental Health Services Research*, 49(2), 237–254. <https://doi.org/10.1007/s10488-021-01160-4>
- Caron, E. B., Muggeo, M. A., Souer, H. R., Pella, J. E., & Ginsburg, G. S. (2020). 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>
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), 284–290. <https://doi.org/10.1037/1040-3590.6.4.284>
- Collyer, H., Eisler, I., & Woolgar, M. (2020). Systematic literature review and meta-analysis of the relationship between adherence, competence and outcome in psychotherapy for children and adolescents. *European Child & Adolescent Psychiatry*, 29(4), 417–431. <https://doi.org/10.1007/s00787-018-1265-2>
- Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, 41(3–4), 327–350. <https://doi.org/10.1007/s10464-008-9165-0>
- Forgatch, M. S., Patterson, G. R., & DeGarmo, D. S. (2005). Evaluating fidelity: Predictive validity for a measure of competent adherence to the Oregon model of parent management training. *Behavior Therapy*, 36(1), 3–13. [https://doi.org/10.1016/S0005-7894\(05\)80049-8](https://doi.org/10.1016/S0005-7894(05)80049-8)
- Greene, R. W., & Ablon, J. S. (2005). *Treating explosive kids: The collaborative problem-solving approach*. Guilford Press.
- 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>
- Hogue, A. (2022). Behavioral intervention fidelity in routine practice: Pragmatism moves to head of the class. *School Mental Health*, 14(1), 103–109. <https://doi.org/10.1007/s12310-021-09488-w>
- Hogue, A., Bobek, M., Porter, N., MacLean, A., Bruynesteyn, L., Jensen-Doss, A., & Henderson, C. (2022). Therapist self-report of fidelity to 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., 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*, 42(2), 229–243. <https://doi.org/10.1007/s10488-014-0548-2>
- Hogue, A., MacLean, A., Bobek, M., Porter, N., Bruynesteyn, L., Jensen-Doss, A., & Henderson, C. E. (2022). Pilot trial of online measurement training and feedback in family therapy for adolescent behavior problems. *Journal of Clinical Child and Adolescent Psychology*, Advance online publication, Apr 06 2022. <https://doi.org/10.1080/15374416.2022.2051529>
- Hogue, A., Porter, N., Bobek, M., MacLean, A., Bruynesteyn, L., Jensen-Doss, A., Dauber, S., & Henderson, C. E. (2022). Online training of community therapists in observational coding of family therapy techniques: Reliability and accuracy. *Administration and Policy in Mental Health and Mental Health Services Research*, 49(1), 139–151. <https://doi.org/10.1007/s10488-021-01152-4>
- 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>
- Kimber, M., Barac, R., & Barwick, M. (2019). Monitoring fidelity to an evidence-based treatment: Practitioner perspectives. *Clinical Social Work Journal*, 47(2), 207–221. <https://doi.org/10.1007/s10615-017-0639-0>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Martin, M., Steele, B., Lachman, J. M., & Gardner, F. (2021). Measures of facilitator competent adherence used in parenting programs and their psychometric properties: A systematic review. *Clinical Child and Family Psychology Review*, 24(4), 834–853. <https://doi.org/10.1007/s10567-021-00350-8>
- McLeod, B. D., Sutherland, K. S., Broda, M., Granger, K. L., Cecilione, J., Cook, C. R., & 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>
- Perepletchikova, F. (2011). On the topic of treatment integrity. *Clinical Psychology: Science and Practice*, 18(2), 148–153. <https://doi.org/10.1111/j.1468-2850.2011.01246.x>
- Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology*, 75(6), 829–841. <https://doi.org/10.1037/0022-006X.75.6.829>
- Pollastri, A. R., Epstein, L. D., Heath, G. H., & Ablon, J. S. (2013). The collaborative problem solving approach: Outcomes across settings. *Harvard Review of Psychiatry*, 21(4), 188–199. <https://doi.org/10.1097/HRP.0b013e3182961017>
- Pollastri, A. R., Wang, L., Eddy, C. J., & Ablon, J. S. (2022). An open trial of collaborative problem solving in a naturalistic outpatient setting. *Clinical Child Psychology and Psychiatry*, 28(2), 512–524. <http://doi.org/10.1177/13591045221094387>
- Pollastri, A. R., Wang, L., Rafferty-Helmer, J. N., Hurley, S., Eddy, C. J., Sisson, J., Thompson, N., & Ablon, J. S. (2022). Development and evaluation of an audio coding system for assessing providers' integrity to collaborative problem solving in youth-service settings. *Professional Psychology: Research and Practice*, 53(6), 640–650. <https://doi.org/10.1037/pro0000476>
- PracticeWise. (2018). *Blue Menu of evidence-based psychosocial interventions for youth*. <https://www.practicewise.com/Community/BlueMenu>.
- Schoemaker, K., Mulder, H., Deković, M., & Matthys, W. (2013). Executive functions in preschool children with externalizing behavior problems: A meta-analysis. *Journal of Abnormal Child Psychology*, 41(3), 457–471. <https://doi.org/10.1007/s10802-012-9684-x>
- Schoenwald, S. K., & Garland, A. F. (2013). A review of treatment adherence measurement methods. *Psychological Assessment*, 25(1), 146–156. <https://doi.org/10.1037/a0029715>
- Schoenwald, S. K., Garland, A. F., Chapman, J. E., Frazier, S. L., Sheidow, A. J., & Southam-Gerow, M. A. (2011). Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research*, 38(1), 32–43. <https://doi.org/10.1007/s10488-010-0321-0>
- Smith, J. D., Dishion, T. J., Brown, K., Ramos, K., Knoble, N. B., Shaw, D. S., & Wilson, M. N. (2016). An experimental study of procedures to enhance ratings of fidelity to an evidence-based family intervention. *Prevention Science*, 17(1), 62–70. <https://doi.org/10.1007/s11121-015-0589-0>
- Stirman, S. W. (2020). Commentary: Challenges and opportunities in the assessment of fidelity and related constructs. *Administration and Policy in Mental Health and Mental Health Services Research*, 47(6), 932–934. <https://doi.org/10.1007/s10488-020-01069-4>
- Wang, L., Pollastri, A. R., Vuijk, P. J., Hill, E. N., Lee, B. A., Samkavitz, A., Braaten, E. B., Ablon, J. S., & Doyle, A. E. (2019). Reliability and validity of the thinking skills inventory, a screening tool for cross-diagnostic skill deficits underlying youth behavioral challenges. *Journal of Psychopathology and Behavioral Assessment*, 41(1), 144–159. <https://doi.org/10.1007/s10862-018-9703-5>
- Zadeh, Z. Y., Im-Bolter, N., & Cohen, N. J. (2007). Social cognition and externalizing psychopathology: An investigation of the mediating role of language. *Journal of Abnormal Child Psychology*, 35(2), 141–152. [doi:10.1007/s10802-006-9052-9](https://doi.org/10.1007/s10802-006-9052-9)