Fidelity Assessment in Community Programs: An Approach to Validating Simplified Methodology

Jessica Suhrheinrich^{1,2} · Kelsey S. Dickson^{2,3} · Neilson Chan⁴ · Janice C. Chan^{2,3} · Tiffany Wang^{2,3} · Aubyn C. Stahmer^{2,5}

Published online: 14 February 2019
© Association for Behavior Analysis International 2019

Abstract

Fidelity to intervention protocol is linked to best outcomes for individuals with autism spectrum disorder (ASD; see Boyd & Corley [Autism 5(4):430–441, 2001]; Pellecchia et al. [J Autism Dev Disord 45(9):2917–2927, 2015]); however, fidelity measurement tools that are both accurate and feasible for community use are often not available. In this paper we explore methods for validated simplification of fidelity assessment procedures toward the goal of increased use in clinical practice. Video recordings ($n = 36$) of therapists working with children with ASD were coded using three variations of fidelity assessment methodology (trial-by-trial, 5-point Likert Scale, and 3-point Likert Scale), and the results were compared for exact agreement, mastery criterion agreement, and overall reliability. The results indicated overall a very high percentage of exact agreement (mean 99.44%, range 94.4–100%) and excellent reliability (mean Krippendorff's alpha [$K\alpha$] 1.0) between the trial-by-trial and 5-point Likert Scale across all components; however, the 3-point method may be viewed as being the more feasible strategy within community programs.

Keywords Intervention fidelity · Community · Reliability

Evidence-based interventions (EBI), or interventions determined to be beneficial based on published research, are critical in promoting desired behavior change and best clinical outcomes

Summary of utility of the work for clinicians and/or researchers of behavior analysis

- To track provider strengths and weaknesses to be better able to focus on specific professional development efforts
- Increasing the feasibility of clinicians using fidelity of implementation tools
- Validating a simpler tool that is reasonable for clinicians to use
- Increasing desired child outcomes by increasing clinician fidelity of implementation

✉ Jessica Suhrheinrich
jsuhrheinrich@sdsu.edu

¹ San Diego State University, 5500 Campanile Drive, San Diego, CA 92182-1170, USA

² The Child Adolescent Services Research Center, 3020 Children's Way, MC 5033, San Diego, CA 92123, USA

³ University of California, San Diego, 9500 Gilman Dr, La Jolla, CA 92093, USA

⁴ Loma Linda University, 11130 Anderson St, Loma Linda, CA 92350, USA

⁵ University of California, Davis MIND Institute, 2825 50th Street, Sacramento, CA 95817, USA

(National Research Council, 2001; Odom, Collet-Klingenberg, Rogers & Hatton, 2010). There has been an increased emphasis on the adoption and implementation of EBI over the past decade, including a particular focus on EBI implementation for youth with autism spectrum disorders (ASD), who represent a clinically challenging population for community providers (Wood, McLeod, Klebanoff, & Brookman-Frazee, 2015). Yet the current literature on child outcomes when EBI are used in community programs is not encouraging, with significantly lower effectiveness estimates than those associated with highly controlled research studies (Henngeler, 2004). In other words, children receiving EBI in schools and clinics are not demonstrating the same outcomes as children receiving EBI through research programs. Although the specific factors affecting these differential outcomes have not been clearly identified, some research suggests that child outcomes are related to variation in treatment integrity or fidelity of intervention (FI; Boyd & Corley, 2001; Pellecchia et al., 2015).

Defined as the degree to which providers are implementing an intervention as intended by developers, FI is a complex, multi-dimensional construct that has received little attention in both research and practice until recently (Nelson, Cordray, Hulleman, Darrow, & Sommer, 2012; Sanetti & Kratochwill, 2009). A review of school-based experimental studies in the *Journal of Applied Behavior analysis* from

1991 to 2005 indicated that only 30% of 152 studies provided treatment integrity/fidelity data (McIntyre, Gresham, DiGennaro & Reed, 2012). Although terminology varies across different scientific fields—for example, FI may also be referred to as “treatment integrity,” “procedural reliability,” and “treatment adherence”—the conceptual meaning is the same. FI encompasses several components and dimensions, including: (1) content, i.e. the steps delivered; (2) competence or quality, i.e., the skill and judgment used during delivery; (3) quantity, i.e. how much of the intervention was delivered; (4) adherence, i.e., the degree to which prescribed and not proscribed procedures are utilized; and (5) differentiation, i.e., features unique to the intervention (Sanetti & Kratochwill, 2009; Schoenwald & Garland, 2013; Schoenwald, Garland, Chapman, Frazier, & Sheidow, 2011). Each of these dimensions together or separately may drive therapeutic outcomes and can be measured together or separately using various methods (e.g., direct observation and coding, indirect questionnaires; Perepletchikova, Treat, & Kazdin, 2007; Schoenwald et al., 2011). For the purposes of this paper, our evaluation of FI focuses on content, quality, quantity, and adherence during EBI delivery. Future research could address additional components of fidelity.

In both research and practice, FI measurement is necessary to demonstrate the relationship between the application of the treatment (independent variable) and its effect on the targeted behavior (dependent variable). Identification of intervention effectiveness stems from rigorous research in which interventions are generally delivered by highly trained clinicians with high levels of FI. In this context, child outcomes are directly tied to FI, with higher fidelity producing better outcomes (Pellecchia et al., 2015; Schoenwald, Sheidow, Letourneau, & Liao, 2003). When intervention delivery is less controlled, the relationship between FI and child outcomes is less clear. Given high rates of adaptations and low rates of sustainability common during EBI delivery (Lau et al., 2017; Stirman et al., 2012), incorporation of fidelity measurement in community clinics and school programs is a critical step toward supporting effective EBI implementation.

The limited information on provider use of EBI for ASD in the community indicates levels of fidelity that are subthreshold to those required in research (Pellecchia et al., 2015; Suhrheinrich et al., 2013). For example, Pellecchia et al. (2015) observed that although teacher FI was directly associated with better child outcomes, teachers still demonstrated low to moderate levels of FI despite considerable training and support. In the dissemination of EBI to real-world community settings, such as school programs, it is likely that the provider’s correct use of intervention strategies is critical for optimizing child outcomes (Durlak & DuPre, 2008; Dusenbury, Brannigan, Falco, & Hansen, 2003).

Although research indicates FI is important for the sustainment of practice and child outcomes, FI is not

consistently monitored in community programs. In fact, preliminary data from our work suggest that fewer than 40% of community supervisors continue to assess FI even after specific and targeted training in an FI measurement tool (Suhrheinrich, Chan, & Dickson, unpublished data, 2016). A recent survey of special education service providers and leaders across the state of California indicated that only 19% report utilizing a formal FI measurement tool to inform delivery of feedback and support to teachers (Suhrheinrich & Dickson, 2017). This value is consistent with broader findings indicating limited evaluation of ongoing intervention use (Stirman et al., 2012). This lack of FI assessment could be related to inadequate resources, such as FI tools that are not appropriate for community program use.

Little is known about how to best support community providers’ measurement of treatment integrity or FI; however, developing tailored processes and resources for providing ongoing FI evaluation may be beneficial (Aarons, Hurlburt, & Horwitz, 2011). The development of fidelity tools involves first identifying important treatment components, or “key ingredients,” developing an instrument that allows for valid and reliable measurement of these components and, in a best case scenario, developing a measure that is psychometrically sound (Schoenwald et al., 2011). However, most current procedures for the measurement of FI come from research studies in which fidelity measurement is relatively complicated. In research settings, FI is often measured using observational methods that involve an observer coding behavior either by live observation or via video review. Such assessors identify, evaluate, and rate the use of key components based on detailed behavioral descriptions that indicate the content, quantity, and/or quality of each component (Mandell et al., 2013; Schoenwald & Garland, 2013; Stahmer et al., 2016). These direct and detailed methods are often considered the gold standard for measuring the appropriate use of intervention strategies. However, in practice, training staff to code observations live in the service setting, and at a similar intensity as is done in research settings, is time-consuming, costly, and possibly not feasible given the time constraints common to community settings (Gearing, El-Bassel, & Ghesquiere, 2011; Perepletchikova et al., 2007; Schoenwald et al., 2011). Thus, the development and introduction of feasible FI tools has the potential to increase regular evaluation of FI.

There have been some recent efforts to increase the evaluation and measurement of FI in community programs, and several research-validated and widely available interventions have begun to incorporate FI tools in their materials for practitioner use, such as the Triple P parenting program (Sanders, Markie-Dadds, & Turner, 2001) and Parent–Child Interaction Therapy (PCIT; Eyberg, 1999). Several ASD-specific interventions also incorporate a fidelity assessment or performance evaluation tool in their materials, including the Early Start Denver Model (Rogers & Dawson, 2010), Parent Training

(Johnson, Handen, & Butter, 2007), and Teaching Social Communication (Ingersoll & Dvortcsak, 2009). The National Professional Development Center (NPDC) Autism Focused Intervention Resources and Modules (AFIRM) provide FI tools that employ a Yes/No coding system (Yes = implemented, No = did not implement) for identified EBI for ASD (AFIRM, 2015). These checklists are useful in guiding both the planning and implementation of EBI. Additionally, some research has involved training supervisory staff within clinical settings to use FI assessment as part of larger efforts toward developing and maintaining effective programs (e.g., Suhrheinrich, 2015). However, the measurement accuracy or reliability of these fidelity tools has not been evaluated. Therefore, in addition to fit and feasibility, inaccurate measurement of intervention delivery may be another factor impacting FI assessment, thereby contributing to the attenuated outcomes seen in community programs.

In promising work, Hogue et al. (2014) evaluated the reliability and validity of a provider-report measure to assess FI for a manualized, family-based preventative intervention. Results from this preliminary work support the reliability and utility of the provider-report checklist for assessing fidelity (Hogue et al., 2014). Additionally, Beidas et al. (2016) explored the reliability of chart-stimulated recall and behavioral rehearsal in evaluating fidelity with positive results. These efforts are encouraging and highlight the potential for researchers to develop, test, and validate the effectiveness of FI measurement tools that fit the needs of community providers.

The purpose of the project reported here was to explore methods for simplification of FI assessment toward the goal of increased use of FI assessment procedures in clinical practice for ASD. For demonstration, one specific multi-component EBI was selected and multiple approaches to FI measurement were compared. Pivotal Response Training (PRT) is a naturalistic, behavioral intervention endorsed by several independent reviews as an EBI for children with autism (Humphries, 2003; National Autism Center, 2009; National Research Council, 2001; Odom, Collet-Klingenberg, & Rogers, 2010; Wong et al., 2015). PRT addresses ‘pivotal’ areas of development, including responsivity to multiple cues, motivation, and independence (Koegel, Koegel, Harrower, & Carter, 1999; Koegel et al., 1989). The targeting of these pivotal areas results in changes in other areas of functioning, thereby reducing the duration of treatment. Implementation of PRT involves a series of prescribed components guiding practitioner behavior. PRT was selected as the focal EBI for this study because although it is widely used in community programs, data suggest practitioners use only some of the components or fail to use all components within the same intervention session (Mandell et al., 2013; Stahmer et al., 2016). Therefore, variable fidelity of the strategies is likely. For this project, we worked toward validation of a simplified PRT FI measure by examining similarities, differences, and reliability in FI measures across three methods

of coding, ranging from extremely rigorous (trial by trial) to highly simplified (3-point [3-Pt] scale).

Method

Procedure

The current project employed three variations of FI assessment methodology to evaluate reliability in coding outcomes using video samples. After the video samples were selected, each video was coded using each of the three coding measures (described in following sections) by trained independent coders. Outcomes and results of each of the three FI measurements were then compared.

Video Samples

Video recordings were drawn from a larger set of videos gathered to examine PRT use in community-based research programs (Stahmer et al., under review). The archived video data were drawn from three separate research trials that involved training providers to use PRT (Jobin, 2012; Schreibman & Stahmer, 2014; Stahmer et al., 2016): (1) a randomized trial including PRT, in which the majority of treatment was provided in-home by trained bachelor’s level and undergraduate student therapists supervised by master’s level Board Certified Behavior Analysts (BCBA; Schreibman & Stahmer, 2014); (2) a single-subject examination of the individualization of PRT in an alternating treatment design that involved undergraduate student therapists implementing PRT in-home, supervised by a master’s level BCBA, (Jobin, 2012); and (3) a study examining the use of PRT in school settings by teachers working in preschool to third grade special education classrooms (Stahmer et al., 2016). The full data set included providers with varied levels of experience and education to ensure a range of FI of PRT, as provider experience and education are known to impact implementation (Aarons, 2004; Lau et al., 2017; Reding, Chorpita, Lau, & Innes-Gomberg, 2014), as are child-level characteristics, including age, gender, race/ethnicity, ASD severity and language level.

From the overall set of 290 usable videos, we randomly selected a subset of 36 videos from across the three archival data sets. The middle 10 min of each session was selected for coding in an attempt to code behavior that reflected how a therapeutic session typically runs, without including “set-up” or “wrap-up” time in which the therapist might be gathering materials, arranging the environment, recording data, or cleaning up.

Participants

Providers Participants included 23 providers trained in PRT strategies as part of clinical research studies (Jobin, 2012;

Schreibman & Stahmer, 2014; Stahmer et al., 2016). All providers were female. Providers included special education teachers ($n = 7$; 30%), undergraduate research assistants ($n = 13$; 57%), and behavior interventionists ($n = 3$; 13%). Please see Table 1 for a complete description of provider participants. Approximately half of providers ($n = 10$) appeared in one video, and 13 appeared in two videos.

Children A total of 19 children who took part in the original research studies (Jobin, 2012; Schreibman & Stahmer, 2014; Stahmer et al., 2016) participated in this study. Child participants included 12 boys (53%) and seven girls (47%), with an average age of 49 (range 18–95) months. The Autism Diagnostic Observation Schedule, 2nd Edition (Lord et al., 2012) was administered to confirm diagnosis. Please see Table 1 for a complete description of child participants. Two children appeared in only one coded video and 17 children appeared in two coded videos.

Coders Coders included 12 research staff and interns with training in PRT. Coders were trained using video samples and coding keys. Coding keys were developed through consensus coding by the intervention developer and experienced clinical supervisors. During consensus coding, clinical supervisors and the intervention developer coded the videos independently. After coding was finished, any discrepancies were discussed, and one code, decided by all members of the group, was assigned. Each coder was trained in only one method of coding. Training continued until the coder independently met an 80% agreement criterion across all behaviors in the coding method over three separate practice videos. Following initial training, inter-rater reliability was examined on an ongoing basis to protect against coder drift. When there were discrepancies between raters, consensus coding was utilized, in which both raters discussed disagreements until one code was decided on.

Inter-Rater Reliability For each coding system, 30% of videos from the sample were randomly selected to be coded by a second coder to allow for analysis of inter-rater reliability. Agreement between the two coders was calculated for each component (Table 2). Overall inter-rater reliability was calculated, with an average Cohen's kappa for trial by trial coding of 0.79 (Range = .066–0.95), an average interclass correlation (ICC) for the 5-point (5-Pt) scale of 0.68 (ICC range .023–0.95) and an average ICC for the 3-Pt scale of 0.42 (ICC range = 0.74–0.94). In terms of significance, Cohen's kappa and ICC are considered to provide good to excellent measurement per current reliability guidelines (Cicchetti, 1994).

Measures

Trial by Trial Coding Trial-by-trial (TBT) coding was considered to be the most rigorous form of FI measurement, requiring

coders to record occurrence/nonoccurrence of each PRT component for every individual opportunity in which the child was expected to respond. TBT coding definitions for ten PRT components were developed with input from experts in PRT in both clinical and research settings using the Delphi method (see Stahmer et al., under review, for a full description of the process). Coders were permitted to rewind and review the video multiple times if needed. Highly specific definitions were used to support coding of each PRT component during each presentation trial in the clip. Coding videos using the TBT method took about 60 min per video, and coders coded approximately ten videos to reach training reliability standards. See Table 3 for a summary of the TBT coding definitions.

5-Point Likert Scale Coding The 5-Pt coding definitions were developed by adapting TBT coding definitions for each PRT component (Table 2; please contact first author for full definitions). For example, language was added to each definition to indicate how often the correct behavior should be observable throughout the session, rather than just during one teaching trial. Anchors were also developed to indicate coding guidelines for each point within the scale. The 5-Pt coding measure included five numerical codes, with associated behavioral definitions and anchors indicating percentage of correct use for each behavior (e.g., a score of 4 indicates that the provider implemented the component correctly 80–99% of the time throughout the session). The coder was instructed to view the full video sample, then to make a coding determination for each PRT component (5 = “Provider implements completely throughout the session” to 1 = “Provider does not implement during the session or never implements appropriately”). Coders were permitted to rewind and review the video multiple times if needed. Permitting a detailed analysis while allowing for appropriate variability in adjusting intervention components based on client behavior, the 5-Pt Likert Scale most closely approximates the FI tools typically utilized in clinical research (Ingersoll & Dvortcsak, 2009; Rogers & Dawson, 2010; Stahmer, Suhrheinrich, Reed, Bolduc, & Schreibman., 2011). See Table 3 for a summary of the 5-Pt coding definitions. Coding videos using the 5-Pt method took approximately 20 min per video, and coders needed to code approximately seven videos to reach training reliability standards.

3-Point Likert Scale Coding The 3-Pt coding definitions were developed by adapting and simplifying the 5-Pt coding definitions for each PRT component. The 3-Pt coding measure included three numerical codes, with associated behavioral descriptions. The coder was instructed to view the full video sample, then make a coding determination for each PRT component (3 = “Provider implements completely throughout the session.” to 1 = “Provider does not implement during the session or never implements appropriately.”). Coders were instructed to review the video once through before providing codes to approximate a

Table 1 Participant demographics

Providers (<i>N</i> = 23)		Children (<i>N</i> = 19)	
Characteristics	Values	Characteristics	Values
Gender		Gender	
Male	0 (0.0%)	Male	12 (63.2%)
Female	23 (100.0%)	Female	7 (36.8%)
Education		Mean age (in months)	47.0 [23.0]
Masters/Doctoral Degree	5 (21.7%)	Race	
Bachelor's Degree/Teaching Credential	4 (17.4%)	White	12 (63.2%)
Associate's Degree	1 (4.3%)	Asian	1 (5.3%)
Current College Student	13 (56.5%)	More than one race	1 (5.3%)
Professional Title		Not reported	5 (26.3%)
Research Assistant	13 (56.5%)	Ethnicity	
Special Education Teacher	7 (30.4%)	Hispanic/Latino	6 (31.6%)
Clinician	3 (13.0%)	Not Hispanic/Latino	8 (42.1%)
Race		Not reported	5 (26.3%)
White	13 (56.5%)	Mean ADOS-2 Comparison Score ^a	7.50 [1.71]; range 4–10
Asian	2 (8.7%)	Receptive Language Age Equivalence Scores (in months)	
American Indian/Alaska Native	2 (8.7%)	Mean MSEL ^b	9.90 [4.93]
Not reported	6 (26.1%)	Mean PLS-4 ^c	23.75 [10.46]
Ethnicity			
Hispanic/Latino	3 (13.0%)		
Not Hispanic/Latino	14 (60.9%)		
Not reported	6 (26.1%)		
Setting			
Home	16 (69.6%)		
Classroom	7 (30.4%)		

Values are presented as a number with the percentage in parenthesis or as the mean with the standard deviation in square brackets

^a Autism Diagnostic Observation Schedule, Second Edition,

^b Mullen Scales of Early Learning

^c Preschool Language Scale, Fourth Edition

live observation. Coders were not permitted to view the video multiple times. No anchors indicating percent of correct use were provided to coders to align with available measures of FI in the community (i.e., NPDC, 2018). See Table 3 for a summary of the 3-Pt coding definitions. Coding videos using the 3-Pt method took about 15 min per video to code.

Analysis

Comparison criteria were developed to examine overall agreement between measures of fidelity. Each numerical code on the 3-Pt and 5-Pt fidelity measures was assigned a corresponding range of percentage of component use from the TBT coding (see Table 4). The specific TBT percentages were selected to best correspond to the coding definitions. Coding outcomes were analyzed across coding systems, and both agreement and reliability were evaluated using several methods. Exact agreement was evaluated by determining the percentage of video

units in which the 3-Pt and 5-Pt code corresponded with the TBT equivalent frequency percentages. Specifically, we examined percentage of exact agreement (e.g., a 5-point Likert Scale rating of 4 and a TBT rating between 80 and 99%). Percentage of agreement regarding meeting mastery criterion for PRT was also evaluated, such that we calculated the percentage of cases in which there was agreement regarding meeting mastery criteria on corresponding rating scales (i.e., a rating of 3 on the 3-Pt Likert Scale, ≥ 4 on the 5-Pt Likert Scale, and 80% frequency or better on the TBT). Finally, Krippendorff's alpha ($K\alpha$) was calculated to evaluate overall reliability between measures for each of the ten components. $K\alpha$ is considered a good index of reliability that is generalizable across differing scales of measurements (such as those used in the current study) as well as robust to missing data (i.e., occasions when certain codes are missing or omitted due inability to code; Hayes & Krippendorff, 2007). $K\alpha$ was calculated using the SPSS KALPHAS Macro (Hayes, 2006). To

Table 2 Coding definitions and reliability for provider behaviors

PRT components	Definitions	Reliability for TBT coding—Cohen's kappa	Reliability for 5-point Likert Scale—ICC	Reliability for 3-point Likert Scale—ICC
Student attention	Child is attending to the provider before the cue is provided either in proximity or orientation toward the provider.	0.77	0.39	– 0.49
Clear cues	Cue should be spoken in clear language or gestural expression.	0.79	0.23	0.77
Developmentally appropriate cues	Cue should be developmentally appropriate and should be provided at the child's or slightly above the child's response level.	0.79	0.86	0.78
Shared control	Provider follows the child's interests and includes preferred materials or activity. Provider moves on to new materials or activity if the child loses interest. Provider takes or facilitates turns while interacting with the child.	0.77	0.78	0.86
Maintenance/acquisition task	Maintenance task: The child correctly responds to the cue 80% of the trials Acquisition task: The child correctly responds to the cue on fewer than 80% of the trials.	0.81	0.41	0.26
Turn taking	Provider takes or facilitates turns while interacting with the child.	0.86	0.86	0.03
Contingent consequence		0.80	0.41	– 0.33
Direct reinforcement	Provider uses contingent, tangible reinforcement for correct behaviors and attempts at correct responding, that is directly related to the teaching activity.	0.81	0.82	0.94
Reinforcement of attempts		0.75	0.86	0.68
Reinforcement of appropriate behavior		0.70	0.95	0.67

ICC Interclass-correlation, PRT Pivotal response training, TBT trial-by-trial,

conduct these analyses, codes were converted to similar metrics (e.g., TBT converted to 5-Pt or 3-Pt codes) utilizing the identified corresponding range of percentage of component

used, as mentioned above (see Table 3). Finally, directional analyses were conducted to examine the nature and direction of discrepancies in coding methods.

Table 3 Coding criteria with descriptive anchors and percentage of use anchors

TBT coding	5-Point Likert Scale	3-Point Likert Scale
Each teaching trial was coded for the presence or absence of provider use of PRT strategies within the trial.	5 “Provider implements competently throughout the session” (100%)	3 “Provider implements competently most of the time, but misses some opportunities”
Frequency data were then aggregated across each minute to facilitate comparison with other coding scales.	4 “Provider implements competently most of the time, but misses some opportunities” (80–99%)	Provider implements competently throughout the session”
	3 “Provider implements competently half the time, but misses many opportunities” (50–79%)	2 “Provider implements competently occasionally, but misses many opportunities”
	2 “Provider implements competently occasionally, but misses many opportunities” (30–49%)	“Provider implements competently half the time, but misses many opportunities”
	1 “Provider does not implement during the session or never implements appropriately” (0–29%)	1 “Provider does not implement during the session or never implements appropriately”
	0 (Not Applicable) “Provider does not have the opportunity to implement during the session”	0 (Not Applicable) “Provider does not have the opportunity to implement during the session”

Comparison range for TBT coding	Rating
5-Point Likert Scale	
100%	5
80–99%	4
50–79%	3
30–49%	2
0–29%	1
	0 (N/A)
3-Point Likert Scale	
67–100%	3
34–66%	2
0–33%	1
	0 (N/A)

Overall the results indicate variable (ranging from low to high) agreement among the TBT, 5-Pt and 3-Pt coding methods across PRT components. Individual comparisons between scales and components are presented in Table 5.

The results indicated overall a very high percentage of exact agreement between the TBT and 5-Pt Likert Scale across all components (mean [M] 99.44%, range 94.4–100%).

The results for the percentage of exact agreement between the converted TBT codes and the 3-Pt Likert ratings indicated variable low to moderate agreement across components, with an average of 66.66% exact agreement (range 44.40–83.30%). Average percentage agreement across components was higher for mastery criteria agreement (M 70.83%, range 63.70–83.30%).

The results indicate variable low to moderate percentage of exact agreement across components between the converted 5-Pt and the 3-Pt Likert scale ratings ($M 58.61\%$, range 27.80–83.30%). Similarly, there was variable moderate agreement across components for meeting mastery criteria ($M 65.83\%$, range 47.20–83.30%).

Table 5 Percentage of videos with agreement between trial-by-trial and Likert Scale coding for individual pivotal response training components

Component	TBT to 5-point LS		TBT to 3-point LS		5-point to 3-point	
	Percentage of exact agreement	Mastery criteria met agreement (%)	Krippendorff's alpha ($K\alpha$)	Percentage of exact agreement	Mastery criteria met agreement (%)	Krippendorff's alpha ($K\alpha$)
Student attention	100.0	100.0	1.0	72.2	72.2	0.09
Clear cue	100.0	100.0	1.0	83.3	83.3	0.29
Developmental appropriate	100.0	100.0	1.0	83.3	83.3	0.30
Shared control	100.0	100.0	1.0	66.7	63.9	-0.19
Maintenance/acquisition	94.4	100.0	.98	50.0	44.4	0.46
Turn taking	100.0	100.0	1.0	55.6	63.9	0.49
Contingent consequences	100.0	100.0	1.0	72.2	55.6	-0.15
Reinforcement of appropriate behavior	100.0	100.0	1.0	66.7	38.9	0.15
Direct reinforcement	100.0	100.0	1.0	72.2	55.6	0.56
Reinforcement of attempts	100.0	100.0	1.0	44.4	61.1	< -0.01

Krippendorff's alpha

Krippendorff's alpha was used to evaluate overall agreement between fidelity measures, including calculating a mean $K\alpha$ across all ten components. The results indicated excellent reliability between the converted TBT ratings and the 5-Pt Likert ($M_{K\alpha}$ 1.0), moderate to low reliability between the TBT ratings and the 3-Pt Likert scale ($M_{K\alpha}$ 0.23), and low reliability between the converted 5-Pt Likert Scale and 3-Pt Likert Scale ($M_{K\alpha}$ 0.18).

Directional Comparisons

Additional analyses were completed to evaluate the nature and direction of disagreements between the three coding methods when they did occur. For example, as indicated in Table 5, there was full agreement between coding methods evaluating the PRT components Maintenance and Acquisition on 94.4% of videos. In the remaining 5.6% of videos, the TBT coding method rendered a higher FI score than the 5-Pt Likert Scale coding method. This analysis allows for a more thorough examination of which coding methodology might be more "lenient" across components and aids interpretation of the outcomes. That is, it permits examination of which coding methods rate provider FI higher or lower compared to other rating methods. For the TBT to 3-Pt comparison, our results indicate that five of the ten PRT components were rated more highly by coders using the TBT method than the Likert Scale method; the remaining five PRT components were equally divided between the TBT versus Likert Scale. For the 5-Pt to 3-Pt comparisons, five of the ten PRT components were rated more highly by coders in the 5-Pt Scale, three components were rated more highly by coders in the 3-Pt Scale and two components were rated roughly equal in both scales. In terms of meeting the mastery criterion, components were rated as meeting the mastery criterion more often using the 5-Pt Scale compared to the 3-Pt Scale, but comparisons between the TBT and 3-Pt Scale indicated very similar component mastery ratings.

Discussion

Evaluation of FI is critical in both intervention development research and for the training and evaluation of community practice to ensure a clear understanding of the independent variable in research studies and service quality in community programs. Toward the goal of increasing the use of FI assessment in both research and community care, this study explored the reliability and agreement of coding outcomes across three FI coding tools with the aim to explore the level of complexity needed to determine treatment integrity and to potentially validate simplified methods of FI coding.

The results from this work lend support for the adaptation of the most rigorous FI assessment methods to formats that are less complex for use in both research and practice. Our examination of agreement between TBT and 5-Pt coding methodologies resulted in high levels of agreement of individual PRT components. This suggests that use of the 5-Pt coding method provides a similar level of accuracy in fidelity measurement as does the TBT coding methods. In addition to being less time consuming, the 5-Pt coding approach is significantly less complex to complete and therefore may require less time to learn. Based on the results presented here, the 5-Pt method represents a more feasible FI measure that supports detailed and accurate measurement of implementation. Compared to the TBT, the Likert Scale method offers a number of advantages for community use, including the limited timing required for use while maintaining accuracy of evaluation. Overall, these results show promise for the use of a simplified scale for the purpose of clinical training in the community, and a simpler system may improve the likelihood of use of any FI assessment.

Whereas our results provide initial support for the utility of a Likert Scale method for FI evaluation, they also suggest that resorting to the simplest method (3-Pt) does not provide the most accurate information on fidelity. Comparison of the TBT and 3-Pt coding methodologies resulted in somewhat lower agreement for several intervention components. Highly varied agreement between the coding methods, as determined by both percentage agreement and Krippendorff's alpha suggest that the 3-Pt coding measure is not as accurate as the other measures. The comparison of 5-Pt and 3-Pt coding methodologies further supports this outcome, with low agreement between measures. Our directional comparisons suggest that 5-Pt method consistently yielded a higher FI score than the 3-Pt method. Similarly, examination of the variability in measures indicate generally greater variability in FI ratings on the 5-Pt scale than on the 3-Pt scale, suggesting that when available, raters utilize the greater response range, thus allowing for more accurate or specific ratings needed for data analysis in research. Together, our findings imply that the 3-Pt measure is not recommended as a reliable research tool to evaluate consistency of PRT implementation. Instead, the more moderate 5-Pt method appears to be both rigorous and practical to support daily community use.

Despite support for the use of the 5-Pt measure, further evaluation regarding the two Likert Scale methods is needed to determine which is more appropriate for community use. This conclusion is supported by the notion that the 3-Pt method more closely approximates existing fidelity forms, including the NPDC fidelity tools (NPDC, 2018), suggesting a perceived utility of this method due to its simplicity. Further, the results of our pass/no pass criterion comparison indicated similar outcomes for all measures, and that the 3-Pt measure was more stringent in terms of evaluating providers' correct use of all components (passing). Consideration of the pass/no pass

criterion is important for measuring FI in community programs because it often drives clinical decision-making around training. For example, in clinical practice, a supervisor may use the pass/no pass criterion to determine if a provider needs additional training before working with clients. Moreover, patterns in FI codes across providers throughout an organization might inform larger training needs and how to best allocate limited resources. For example, if multiple providers show weakness in implementation of one or more components, these might be selected as the focus of professional development efforts.

Currently, however, little is known on fidelity assessment and measurement in the community, including which measures are viewed as acceptable for these settings. It is possible that the 5- and 3-Pt methods are both feasible tools for assessing fidelity in some community practices, and there may be value in exploring additional modifications to training methodology and the coding anchors; as well, expanded definitions of the components may be necessary to create a simpler system that more accurately matches more rigorous coding methods. For example, it is possible that the higher reliability of the 5-Pt method stems from the greater response range, enabling a higher level of specificity in coding behavioral definitions that better fit with behavioral providers' prior experience; with further specification and training or support in applying coding definitions, higher reliability for the 5-Pt and even the 3-Pt methods may be possible. Additionally, since coding was completed by researchers, we do not know the validity of the measure when used by community providers. While validation of simplified FI coding tools is necessary, this alone is likely not sufficient for integration of FI assessment into community practice. There is significant need for ongoing development and targeted integration of feasible and accurate fidelity measures into community program settings. Testing the use of these methods in both research and community programs with ample training and support for FI assessment would greatly inform efforts to increase community FI evaluation. Therefore, an exploration of current use of FI methods in community ASD service programs to gain a better understanding of what methodology is viewed as feasible is needed.

There are several limitations to the current project. First, due to the nature of the study, coders for the current project were trained by the research team and were undergraduate or BA-level research assistants. Although they were provided ample training in the coding process and reached reliability standards prior to coding independently, they had minimal clinical training as part of this research experience. It is possible that clinical practitioners with more experience implementing EBI and working with individuals with ASD will apply the coding methodologies and behavioral definitions differently. Further, there is no direct involvement of clinicians or providers as coders in the current study, which limits our ability to directly speak to the appropriateness of our FI tools among community providers. We recommend clinician coding be integrated into further

research. Methodologically, one limitation relates to coder viewing of the video samples. For 3-Pt coding, each video could only be viewed one time to approximate live coding conditions. However, for TBT and 5-Pt coding, each video could be viewed more than once. Future research should include standard coding protocol across coding systems. Another limitation is the focus on only one EBI for ASD. These findings may not generalize to other interventions for ASD or more broadly. However, the model used for evaluation of FI assessment methodology (i.e., comparison of three FI assessment methods) may be useful in improving feasibility and informing the development and use of FI tools for other interventions. Finally, some components have variable inter-rater reliability across the three rating systems, which may have impacted our comparisons across methodologies. Therefore, some degree of caution is warranted in the interpretation and generalization of our results given that a high degree of inter-rater reliability may be a prerequisite for conducting comparisons between methodologies.

The current project addresses one barrier to evaluation of FI, the complexity of scoring. Moving forward, methods to support sustainment of FI measurement in clinical practice, should also be explored. Evaluating FI throughout intervention delivery is important for ongoing practice sustainment and to determine if additional training and support are needed. Per traditional research methodology, not only do staff rate FI during the initial intervention training, they need to consistently monitor FI throughout the implementation of the intervention to prevent against drift and assure best clinical outcomes (Cooper, Heron, & Heward, 2007; Gresham & Gansle, 1993). In combination with efforts to further develop reliable and practical FI measures, this may require allocation of further supports (e.g., allocated staff time, rewards for the use of FI tools) or other resources for assessment of FI in practice. Policy changes may support integration of FI assessment into regular practice. For example, the Los Angeles County Department of Mental Health launched a Prevention and Early Intervention Transformation (PEI) initiative in 2010 that mandated the use of EBI, including use of FI or performance monitoring strategies, which significantly increased provider's measurement of fidelity (Los Angeles County Department of Mental Health, 2010). However, this level of use likely does not reflect general use of FI measurement when it is not specifically included in training or required.

The importance of FI measurement in the sustainment of effective EBI use necessitates the creation and adoption of measures that balance effectiveness and efficiency (Schoenwald et al., 2011). In other words, the measure and evaluation process should be feasible to use, contain a system for offering or obtaining performance feedback based on the measure, and include a clear link between FI and child outcomes. The current study represents an important first step, but there is still much work to be done to achieve the integration of effective FI evaluation in community programs.

Acknowledgements Drs. Stahmer and Suhrheinrich and Mr. Chan were affiliated with the University of California, San Diego Department of Psychiatry at the time this work was completed. Drs. Stahmer and Suhrheinrich are also investigators with the Implementation Research Institute (IRI), at the George Warren Brown School of Social Work, Washington University in St. Louis, through an award from the NIMH (5R25MH08091607).

Funding This work was conducted at the Child and Adolescent Services Research Center and was supported by a National Institute of Mental Health (NIMH) Research Grant (4R21/33MH097033) and Career Development grant (K01MH109574).

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Informed consent was obtained from all individual participants included in the study.

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

References

- Aarons, G. A. (2004). Mental health provider attitudes toward adoption of evidence-based practice: The evidence-based practice attitude scale (EBPAS). *Mental Health Services Research*, 6(2), 61–74. <https://doi.org/10.1023/B:MHSR.0000024351.12294.65>.
- Aarons, G. A., Hurlburt, M., & Horwitz, S. M. (2011). Advancing a conceptual model of evidence-based practice implementation in public service sectors. *Administration and Policy in Mental Health and Mental Health Services Research*, 38(1), 4–23. <https://doi.org/10.1007/s10488-010-0327-7>.
- Autism Focused Intervention Resources & Modules (AFIRM). (2015). *AFIRM online learning modules*. Chapel Hill, NC: National Professional Development Center on Autism Spectrum Disorder.
- Beidas, R., Maclean, J., Fishman, J., Dorsey, S., Schoenwald, S., Mandell, D., et al. (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), 323. <https://doi.org/10.1186/s12888-016-1034-z>.
- Boyd, R. D., & Corley, M. J. (2001). Outcome survey of early intensive behavioral intervention for young children with autism in a community setting. *Autism*, 5(4), 430–441. <https://doi.org/10.1177/1362361301005004007>.
- 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.
- Cooper, J. O. J., Heron, T. E., & Heward, W. L. (2007). Promoting generalized behavior change. In *Applied behavior analysis 2nd Ed* (pp. 613–656). Retrieved from https://scholar.google.com/scholar?q=cooper+Promoting+generalized+behavior+change.+&btnG=&hl=en&as_sdt=0%2C5.
- 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>.
- Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research*, 18(2), 237–256. <https://doi.org/10.1093/her/18.2.237>.
- Eyberg, S. (1999). *Parent–child interaction therapy: Integrity checklists and session materials*. Unpublished manuscript. Gainesville: University of Florida. Retrieved from: https://scholar.google.com/scholar_lookup?hl=en&publication_year=1999&author=S.+M.+Eyberg&title=+Parent%E2%80%93child+interaction+therapy%3A+Integrity+checklists+and+session+materials+&.
- Gearing, R., El-Bassel, N., & Ghesquiere, A. (2011). Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. *Clinical Psychology Review*, 31(1), 79–88. <https://doi.org/10.1016/j.cpr.2010.09.007>.
- Gresham, F., & Gansle, K. (1993). Treatment integrity in applied behavior analysis with children. *Journal of Applied Behavior Analysis*, 26(2), 257–263. <https://doi.org/10.1901/jaba.1993.26-257>.
- Hayes, A. (2006). SPSS macro for computing Krippendorff's alpha. Retrieved from: <http://www.comm.ohio-state.edu/ahayes/SPSS%20programs/kalpha.htm>
- Hayes, A. F. A., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1), 77–89. <https://doi.org/10.1080/19312450709336664>.
- Henngeler, S. W. (2004). Decreasing effect sizes for effectiveness studies—implications for the transport of evidence-based treatments: Comment on Curtis, Ronan, and Borduin (2004). *Journal of Family Psychology*, 18(3), 420–423. <https://doi.org/10.1037/0893-3200.18.3.420>.
- 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>.
- Humphries, B. (2003). What else counts as evidence in evidence-based social work? *Social Work Education*, 22(1), 81–91. Retrieved from: <https://doi.org/10.1080/02615470309130>.
- Ingersoll, B., & Dvortcsak, A. (2009). *Teaching social communication to children with autism: A practitioner's guide to parent training and a manual for parents*. New York, NY: Guilford Press.
- Jobin, A. B. (2012). *Integrating treatment strategies for children with autism*. PhD thesis. Dan Diego: UC San Diego.
- Johnson, C., Handen, B. L., Butter, E., Wagner, A., Mulick, J., Sukhodolsky, D. G., et al. (2007). Development of a parent training program for children with pervasive developmental disorders. *Behavioral Interventions*, 22(3), 201–221. <https://doi.org/10.1002/bin.237>.
- Koegel, L. K., Koegel, R. L., Harrower, J. K., & Carter, C. M. (1999). Pivotal response intervention I: Overview of approach. *Journal of the Association for Persons with Severe Handicaps*, 24(3), 174–185. <https://doi.org/10.2511/rpsd.24.3.174>.
- Koegel, R. L., Schreibman, L., Good, A., Cerniglia, L., Murphy, C., & Koegel, L. K. (1989). *How to teach pivotal behaviors to autistic children*. Santa Barbara, CA: University of California, Santa Barbara.
- Lau, A., Stadnick, N., Regan, J., Roesch, S., Barnett, M., Saifan, D., et al. (2017). Therapist report of adaptations to delivery of evidence-based practices within a system-driven reform of publicly funded children's mental health services. *Journal of Consulting & Clinical Psychology*, 85(7), 664–675. <https://doi.org/10.1037/ccp0000215>.
- Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. L. (2012). *Autism diagnostic observation schedule—2nd edition (ADOS-2)*. Torrance, CA: Western Psychological Services.
- Los Angeles County Department of Mental Health. (2010). *Prevention and Early Intervention Plan*. Los Angeles, CA: Los Angeles County Department of Mental Health.

- Mandell, D. S., Stahmer, A. C., Shin, S., Xie, M., Reisinger, E., & Marcus, S. C. (2013). The role of treatment fidelity on outcomes during a randomized field trial of an autism intervention. *Autism, 17*(3), 281–295. <https://doi.org/10.1177/1362361312473666>.
- McIntyre, L. L., Gresham, F. M., DiGennaro, F. D., & Reed, D. D. (2012). Treatment integrity of school-based interventions with children in the journal of applied behavior analysis 1991–2005. *Journal of Applied Behavior Analysis, 40*(4). <https://doi.org/10.1901/jaba.2007.659-672>.
- National Autism Center. (2009). *National standards report*. Randolph, MA: National Autism Center.
- National Professional Development Center (NPDC, 2018). Autism focused intervention resources and modules. Retrieved from <https://afirm.fpg.unc.edu>.
- National Research Council. (2001). Educating children with autism. Committee on educational interventions for children with autism. In: C. Lord & J. P. McGee (Eds.), Division of behavioral and social sciences and education. Washington, DC: National Academy Press.
- Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *The Journal of Behavioral Health Services and Research, 39*(4), 374–396. <https://doi.org/10.1007/s11414-012-9295-x>.
- Odom, S., Collet-Klingenberg, L., Rogers, S., & Hatton, D. D. (2010). Evidence-based practices in interventions for children and youth with autism spectrum disorders. *Preventing school failure, 54*(4), 275–282. <https://doi.org/10.1080/10459881003785506>.
- Pellecchia, M., Connell, J. E., Beidas, R. S., Xie, M., Marcus, S. C., & Mandell, D. S. (2015). Dismantling the active ingredients of an intervention for children with autism. *Journal of Autism & Developmental Disorders, 45*(9), 2917–2927. <https://doi.org/10.1007/s10803-015-2455-0>. Dismantling.
- 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>.
- Reding, M. E. J., Chorpita, B. F., Lau, A. S., & Innes-Gomberg, D. (2014). Providers' attitudes toward evidence-based practices: Is it just about providers, or do practices matter, too? *Administration and Policy in Mental Health and Mental Health Services Research, 41*(6), 767–776. <https://doi.org/10.1007/s10488-013-0525-1>.
- Rogers, S., & Dawson, G. (2010). *Early start Denver model for young children with autism: Promoting language, learning, and engagement*. New York, NY: Guilford Press.
- Sanders, M. R., Markie-Dadds, C., & Turner, K. M. T. (2001). *Practitioner's manual for standard triple P*. Birsbaine, Australia: Families International Publishing.
- Sanetti, L. M. H., & Kratochwill, T. R. (2009). Toward developing a science of treatment integrity: Introduction to the special series. *School Psychology Review, 38*(4), 445–459.
- 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. (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>.
- Schoenwald, S. K., Sheidow, A. J., Letourneau, E. J., & Liao, J. G. (2003). Transportability of multisystemic therapy: Evidence for multilevel influences. *Mental Health Services Research, 5*(4), 223–239.
- Schreibman, L., & Stahmer, A. (2014). A randomized trial comparison of the effects of verbal and pictorial naturalistic communication strategies on spoken language for young children with autism. *Journal of Autism and Developmental Disorders, 44*(5), 1244–1251. <https://doi.org/10.1007/s10803-013-1972-y>.
- Stahmer, A. C., Suhrheinrich, J., & Rieth, S. (2016). A pilot examination of the adapted protocol for classroom pivotal response teaching. *Journal of the American Academy of Special Education Professionals, 119*–139. Retrieved from <http://search.proquest.com/docview/1826535643?accountid=13963%5Cn>, http://resolver.ebscohost.com/openurl?ctx_ver=Z39.88-2004&ctx_enc=info:ofi/enc:UTF-8&rft_id=info:sid/ProQ%3Aericshell&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&rft.genre=article&rft.title=Gr.
- Stahmer, A. C., Suhrheinrich, J., Roesch, S., Zeedyk, S., Wang, T., Chan, N., & Loo, H. S. (under review). Examining relationships between child skills and potential key components of an evidence-based practice in ASD.
- Stahmer, A., Suhrheinrich, J., Reed, S., Bolduc, C., & Schreibman, L. (2011). *Classroom pivotal response teaching: A guide to effective implementation*. New York, NY: Guilford Press.
- Stirman, S. W., Kimberly, J., Cook, N., Calloway, A., Castro, F., & Charns, M. (2012). The sustainability of new programs and innovations: A review of the empirical literature and recommendations for future research. *Implementation Science, 7*(1), 17. <https://doi.org/10.1186/1748-5908-7-17>.
- Suhrheinrich, J. (2015). A sustainable model for training teachers to use pivotal response training. *Autism, 19*(6), 713–723. <https://doi.org/10.1177/1362361314552200>.
- Suhrheinrich, J., & Dickson, K. S. (2017). Mapping leadership structure in special education programs to tailor leadership intervention. In *The Society for Implementation Research Collaboration*. Seattle, WA.
- Suhrheinrich, J., Stahmer, A. C., Reed, S., Schreibman, L., Reisinger, E., & Mandell, D. (2013). Implementation challenges in translating pivotal response training into community settings. *Journal of Autism & Developmental Disorders, 43*(12), 2970–2976. <https://doi.org/10.1007/s10803-013-1826-7>.
- Wong, C., Odom, S. L., Hume, K. A., Cox, A. W., Fetting, A., Kucharczyk, S., et al. (2015). Evidence-based practices for children, youth, and young adults with autism spectrum disorder: A comprehensive review. *Journal of Autism and Developmental Disorders, 45*(7), 1951–1966. <https://doi.org/10.1007/s10803-014-2351-z>.
- Wood, J. J., McLeod, B. D., Klebanoff, S., & Brookman-Frazee, L. (2015). Toward the implementation of evidence-based interventions for youth with autism spectrum disorders in schools and community agencies. *Behavior Therapy, 46*(1), 83–95. <https://doi.org/10.1016/j.beth.2014.07.003>.