Title

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

Journal

Behavior Analysis in Practice, 13(1)

ISSN

1998-1929

Authors

Suhrheinrich, Jessica
Dickson, Kelsey S
Chan, Neilson
et al.

Publication Date

2020-03-01

DOI

10.1007/s40617-019-00337-6

Peer reviewed

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

1. Jessica Suhrheinrich, Ph.D., San Diego State University and the Child and Adolescent Services Research Center 5500 Campanile Drive, San Diego, CA 92182-1170 USA. Email: jsuhrheinrich@sdsu.edu

2. Kelsey S. Dickson, Ph.D., University of California, San Diego and the Child and Adolescent Services Research Center, 3020 Children's Way MC5033, San Diego, CA, 92123, USA; email: ksdickson@ucsd.edu

3. Neilson Chan, M.A., Loma Linda University, 11130 Anderson St, Loma Linda, CA 92350, USA; email: nechan@llu.edu

4. Janice C. Chan, M.A., BCBA, University of California, San Diego and the Child and Adolescent Services Research Center, 3020 Children's Way MC5033, San Diego, CA, 92123, USA; email: jdc012@ucsd.edu

5. Tiffany Wang, B.S., University of California, San Diego and the Child and Adolescent Services Research Center, 9500 Gilman Dr. La Jolla CA 92093, USA; email: t8wang@ucsd.edu

6. Aubyn C. Stahmer, Ph.D., University of California, Davis MIND Institute and the Child and Adolescent Services Research Center, 2825 50th Street, Sacramento, CA 95817, USA; email: astahmer@ucdavis.edu

Acknowledgements

Please note that 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. This work was conducted at the Child and Adolescent Services Research Center. This research was supported by a National Institute of Mental Health (NIMH) Research Grant (4R21/33MH097033) and Career Development grant (K01MH109574). Additionally, Drs. Stahmer and Suhrheinrich are 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).The authors have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Jessica Suhrheinrich, Ph.D. Department of Special Education, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182-1170 USA. Email: jsuhrheinrich@sdsu.edu

Keywords: Intervention Fidelity, Community-Use, Reliability

As demand for autism intervention services increases, it is critical that community agencies effectively implement evidence-based interventions (EBI), or interventions which research has determined are beneficial for children with autism spectrum disorders (ASD). Recent efforts to implement EBI for autism have been influenced by practice reviews including: The National Standards Project (NSP), which identified 11 categories of interventions as "established," and the National Professional Development Center (NPDC), which identified 27 focused intervention practices for ASD known to have positive outcomes (National Autism Center, 2009; NPDC, 2010, Wong et al, 2015). These independent reviews had significant overlap in their respective findings, indicating strong support for specific interventions and their ability to address the symptoms and needs of individuals with ASD.

Both the NSP and NPDC reports have informed significant efforts towards dissemination and scale-up of EBI for ASD. In 2010, NPDC launched a multistate comprehensive professional development process aiming to promote teacher and provider use of EBI (NPDC, 2010) which is currently being updated based on more recent additions to the EBI literature. Additionally, there have been an increasing number of community-based training studies aimed at increasing teacher and provider training in EBIs. These have been conducted in multiple settings including (1) schools where researchers have conducted two randomized trials after training teachers in the Strategies for Teaching based on Autism Research (STAR) program (e.g., Mandell et al., 2013) and an adaptation of Pivotal Response Training specifically for classroom use, Classroom Pivotal Response Teaching (CPRT; Stahmer, Suhrheinrich, & Rieth, 2016); (2) early intervention settings, including training early intervention providers to use a parent-implemented intervention, Project ImPACT, in homes (Stahmer, Rieth, Stoner, Feder, Searcy, & Wang, 2017; Stadnick, Stahmer & Brookman-Frazee, 2015) as well as (3) mental health settings using the

Individualized Mental Health Intervention for Children with ASD (AIM HI) model (Brookman-Frazee, Drahota & Stadnick, 2012). Despite the increasing commitment to increasing provider training in EBI, we continue to have limited understanding of how well these providers are implementing these EBI in the community once the research support has ended.

Accurate implementation and sustainment are important because the literature on child outcomes when EBI are implemented in community programs is not encouraging, with significantly lower effectiveness estimates when interventions are compared to RCTs (Hennggeler, 2004). Though the specific factors affecting these differential outcomes have not been clearly identified, some research suggests that differences may be intricately tied to variation fidelity to the intervention, or how well providers are implementing the intervention strategies (Boyd & Corley, 2001; Pellecchia et al., 2015). Fidelity of intervention (FI) is the degree to which an intervention is implemented as it was intended by the developers (Nelson, Cordray, Hulleman, Darrow, & Sommer, 2012). 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 child behavior (dependent variable). Our current understanding regarding the effect of an intervention stems from rigorous RCTs 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). Unfortunately, the limited information on provider use of EBI for ASD in the community indicate levels of fidelity that are subthreshold to those required in research (Pellecchia et al., 2015; Suhrheinrich et al., 2013). For example, Pellecchia and colleagues (2015) observed that despite considerable training and support surrounding implementation, teachers demonstrated limited FI during delivery and teacher's FI

was directly associated with better child outcomes. In the dissemination of EBI to community settings, it is likely that the provider's correct use of intervention strategies is critical for the optimal benefit of the child (Durlak & DuPre, 2008; Dusenbury, Brannigan, Falco, & Hansen, 2003).

One way to improve the use and sustainment of EBI in community settings may be incorporation of systems or processes for providing ongoing FI evaluation (Aarons, Hurlburt, & Horwitz, 2011). However, current measurement of FI comes from research studies where fidelity measurement is relatively complicated. 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). Conceptually, FI may include the occurrence (whether or not a behavior occurs), frequency, and/or quality of the key ingredients (Schoenwald & Garland, 2013). In research settings, FI is often measured using observational methods that involve an observer coding behavior either by live observation or via video review. Assessors identify, evaluate and rate the use of key components based on detailed descriptions of the prescribed components indicating occurrence, frequency 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 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, however, is potentially time-consuming, costly and not feasible given time constraints common to community settings (Gearing, El-Bassel, & Ghesquiere, 2011; Perepletchikova, Treat, & Kazdin, 2007; Schoenwald et al., 2011)

Although research suggests FI is important for sustainment and child outcomes, there are many challenges to measuring FI in community settings. In fact, preliminary data from our work suggest that less than 40% of community supervisors continue to assess FI even after specific and targeted training in an FI measurement tool. A recent survey of special education service providers and leaders across the state of California indicates that only 19% report utilizing a formal FI measurement tools to inform delivery of feedback and support to teachers (Suhrheinrich & Dickson, 2017). This lack of FI assessment could be related to inadequate resources, as existing FI tools may not be feasible for community program use.

To address this concern, there have been some recent efforts to increase evaluation and measurement of FI in community programs. For example, several research-validated and widely available interventions have begun to incorporate tools for assessing fidelity in their materials for practitioner use (e.g., Triple P [Sanders, Markie-Dadds, & Turner, 2001], 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 NPDC Autism Focused Intervention Resources and Modules (AFIRM) provide FI assessment tools that employ a Yes/No coding system (Yes=implemented, No=did not implement) for identified EBI for ASD (AFIRM Team, 2015). These checklists are useful in guiding both 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 of these fidelity tools has not been evaluated. Therefore, in the case of FI measurement, fit, feasibility, and accuracy are likely significant barriers to use.

In promising work, Hogue and colleagues (2014) created 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 a therapist-report checklist for assessing fidelity (Hogue et al., 2014). Additionally, Beidas and colleagues (2016) report plans to further the development of accurate and feasible fidelity measurement tools for community-delivered Cognitive Behavioral Therapy by exploring the role of chart stimulated-recall and behavioral rehearsal. 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 current project was to explore methods for validated 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; National Standards Report, 2009; 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, variability in FI of the strategies is likely. For this project, we work 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 scale).

Methods

Procedure

The current project employed three variations of FI assessment methodology to evaluate reliability in coding outcomes using video samples. After video samples were selected, each video was coded using each of the three coding measures (described below) 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); (3)study examining the use of PRT in school settings by teachers working in preschool to 3rd

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 is known to impact implementation (Aarons, 2004; Lau et al., 2017; Reding, Chorpita, Lau, & Innes-Gomberg, 2014) and a range of child level characteristics.

From the overall set of 290 usable video, a subset of 36 videos from across the three archival data sets were randomly selected. The middle ten minutes 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; 56.52%) and community clinicians (n=3; 13%). Please see *Table 1* for a complete description of provider participants. About half of providers (10) appeared in one video and 13 appeared in two videos.

Children. Participants included 19 children who took part in the original research studies (Jobin, 2012; Schreibman & Stahmer, 2014; A. Stahmer et al., 2016). Child participants included 10 boys (53%) and 7 girls (47%), with an average age of 49 months (range 18 – 95 months). The Autism Diagnostic Observation Schedule -2 (ADOS-2; 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 1 video and 17 children appeared in two coded videos.

Coders. Coders included twelve research staff and interns with training in PRT. Coders were trained using gold-standard coding keys on the coding methods discussed below. 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, interrater reliability was examined on an ongoing basis to protect against coder drift. When there were discrepancies between raters, consensus coding was utilized.

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 .79 (Range = .66 - .95), an average interclass correlation (ICC) for 5 Pt scale of 0.68 (ICC Range = .23 - .95) and an average ICC for the 3 Pt scale of 0.42 (ICC Range = -.74 - 0.94).

Measures

Trial by Trial Coding. Trial by Trial (TBT) coding was considered 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 10 PRT components were developed with input from experts in PRT in both clinical and research settings using the Delphi method (see Stahmer et al., in 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

minutes per video and coders needed to code 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 point (5 Pt) coding definitions were developed by adapting TBT coding definitions for each PRT component. 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, and anchors were 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 percent of correct use for each behavior. The coder is instructed to view the full video sample, then 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; A. Stahmer et al., 2011). See *Table 3* for a summary of the 5 Pt coding definitions. Coding videos using the 5 Pt method took about 20 minutes per video and coders needed to code approximately seven videos to reach training reliability standards.

3 Point Scale Coding. The 3 point (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 is 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 live observation. No anchors indicating percent of correct use were provided to coders to align with evaluation methodologies common in clinical practice. The 3 Pt Likert scale most closely approximates available and/or feasible measure of FI in the community (i.e., NPDC, 2010). See *Table 3* for a summary of the 3 Pt coding definitions. Coding videos using the 3 Pt method took about 15 minutes per video to code.

Analysis

To examine overall agreement between measures of fidelity, comparison criteria were developed. Each numerical code on the 3Pt and 5Pt fidelity measures was assigned a corresponding range of percent 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 percent of video units in which the 3Pt and 5Pt 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 80-99%). Percentage of agreement regarding meeting mastery criterion for PRT was also evaluated, such that we calculated the percent of cases in which there was agreement regarding meeting mastery criteria on corresponding rating scales (i.e., a rating of three on the 3-point Likert, four or better on the 5-point Likert, and 80% frequency or better on the TBT). Finally, Krippendorff's Alpha (Kα) was calculated to evaluate overall reliability between measures for each of the 10 components. Kα is considered a good index of reliability that is generalizable across scales of measurements (such as those used in the

current study) as well as robust to missing data (A. F. Hayes & Krippendorff, 2007). Kα was calculated using the SPSS KALPHAS Macro (Hayes, 2006). To 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 percent of component used mentioned above (see *Table 3*).

Results

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

TBT to 5 Pt Likert Scale

Results indicated overall a very high percentage of exact agreement between the TBT and 5 Pt scale across all components (M = 99.44%, Range 94.4-100%).

TBT to 3 Pt Likert Scale

The results for the percent 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 percent agreement across components was higher for mastery criteria agreement (M = 70.83%, Range 63.70-83.30%).

5 Pt to 3 Pt Likert Scale

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%).

Krippendorf's Alpha

Krippendorff's Alpha (Kα) was used to evaluate overall agreement between fidelity measures, including calculating a mean Kα across all 10 components. Results indicated excellent reliability between the converted TBT ratings and 5 Pt Likert ($M_{K\alpha}$ = 1.0), moderate to low reliability between the TBT ratings and the 3 Pt Likert scale ($M_{K\alpha}$ = .23) and low reliability between the converted 5 Pt Likert Scale and 3 Pt Likert Scale ($M_{K\alpha}$ =.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 above, there was full agreement between coding methods evaluating 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 more thorough analysis of which coding methodology might be more "lenient" across components and aids interpretation of the outcomes. For the TBT to 3 Pt comparison, results indicate that five of the 10 PRT components are rated more highly by coders using the TBT 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 10 PRT components are rated more highly by coders in the 5 Pt Likert Scale, three components rated more highly by coders in the 3 Pt Likert Scale and two are roughly equal. In terms of meeting mastery criterion, components were rated as meeting mastery criterion more often using the 5 Pt Likert Scale compared to the 3 Pt Likert Scale. In terms of a TBT and 3 Pt Likert Scale comparison, both scales rated components as meeting mastery criterion more frequently roughly an equal amount of time.

Discussion

Evaluation of FI is critical in both intervention development research and for training and evaluation of community practice to ensure 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 reliability and agreement of coding outcomes across three FI coding tools in order to explore the level of complexity needed to determine treatment integrity, and to potentially validate simplified methods of FI coding.

Results from this work lend support for adaptation of the most rigorous FI assessment methods to be less complex for use in both research and practice. Our examination of agreement between TBT and 5Pt coding methodologies resulted in high levels of agreement of individual PRT components. This suggests that use of the 5Pt coding method provides a similar level of accuracy in fidelity measurement as does the TBT coding methods. The 5pt 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 is likely a feasible FI measure that supports detailed, nuanced, and accurate measurement of implementation.

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 percent agreement and Krippendorf's Alpha suggest 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 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.

The pass/no pass criterion comparison for all measures resulted in similar outcomes. Again, the TBT and 5 Pt measures showed strong agreement on which providers met mastery criteria for PRT whereas the TBT and 3 Pt measures had low agreement. Further, the 5 Pt measure showed consistently higher rates of meeting mastery criterion compared to the 3 Pt measure. That is, the 3 Pt measure was more stringent in terms of evaluating providers' correct use of all components (passing). Consideration of 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 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.

Despite low agreement on a component by component level with the other two methods, it is likely that the 3 Pt method may be viewed as more feasible within community programs due to the simplicity of the form, behavioral definitions and coding options. This is supported by notion that the 3 Pt measure more closely approximates existing fidelity forms, including the NPDC fidelity tools (NPDC, 2010). The 3pt method does provide a stringent rating of overall use of the CPRT protocol, which may support its use for sustainment of practice over time. Thus,

these results show promise for using a simplified 3pt scale for the purpose of clinical training in the community and a simpler system may improve the likelihood of use of any FI assessment.

Overall, these outcomes support the use of the 5 Pt measure as an accurate research measure of FI that is comparable to the TBT methodology and that the 3 Pt method may be more appropriate for community use due to its simplicity. However, additional modifications to the coding anchors and/or expanded definitions of the components may be necessary to increase reliability of the simpler system with more detained ones. Additionally, since coding was completed by researchers, we do not know the validity of the measure when used by community providers. Currently, however, little is known regarding 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. 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.

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. Further, 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. Further, there may be value in exploring training methodology or modification of coding definitions that may support use of the 3 Pt scale that more accurately matches more rigorous coding methods.

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 or interpret the behavioral definitions differently. This limits our ability to directly speak to the appropriateness of our FI tools among community providers. 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 for evaluation of FI assessment methodology may be useful in improving feasibility and informing us of FI tools for other interventions.

The current project addresses one barrier to evaluation of FI, the complexity of scoring. However, in addition to unavailability of FI measurement tools, additional barriers to collecting FI data in community settings exist. Additionally, 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 practice, this may require allocation of staff time or other resources for assessment of FI. 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 necessity of a FI measurement tool for sustainment of effective EBI use necessitates the creation and adoption of a measures that balance effectiveness and efficiency (Schoenwald et al., 2011). That is, 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. Thus, the current study represents an important first step but there is still much work to be done in integration of effective FI evaluation in community programs.

References

Aarons, G. (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.

Aarons, G., 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.

AFIRM Team. (2015). AFIRM Online Learning Modules. Chapel Hill, NC: National Professional Development Center on Autism Spectrum Disorder. Retrieved from http://afirm.fpg.unc.edu/

Arick, J. R., Loos, L., Falco, R., & Krug, D. (2004). The STAR program: Strategies for teaching based on Autism research. Retrieved from http://starautismsupport.com/curriculum/star-program

Beidas, R., Maclean, J., Fishman, J., Dorsey, S., Schoenwald, S., Mandell, D., … Marcus, S. C. (2016). A randomized trial to identify accurate and cost-effective fidelity measurement methods for cognitive-behavioral therapy: project FACTS study protocol. *BMC Psychiatry*, *16*, 1–10. https://doi.org/10.1186/s12888-016-1034-z

Boyd, R., & Corley, M. (2001). Outcome survey of early intensive behavioral intervention for young children with autism in a community setting. *Autism*.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). Promoting generalized behavior change. In *Applied behavior analysis 2nd Ed* (pp. 613–656).

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*. 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 Educ Res*, *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. *Retrieved April*.

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*.

Gresham, F., & Gansle, K. (1993). Treatment integrity in applied behavior analysis with children. *Journal of Applied*.

Hayes, A. (2006). SPSS macro for computing Krippendorff's alpha. *Retrieved September*.

Hayes, A. F., & 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

Hennggeler, 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.

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*.

Ingersoll, B., & Dvortcsak, A. (2009). *Teaching Social Communication to Children With Autism: A Practitioner's Guide to Parent Training and a Manual for Parents*.

Jobin, A. B. (2012). *Integrating treatment strategies for children with autism*. *Psychology*. UC San Diego. Retrieved from https://escholarship.org/uc/item/35n2x2kp

Johnson, C., Handen, B., & Butter, E. (2007). Development of a parent training program for children with pervasive developmental disorders. *Behavioral*.

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 children with autism: A training manual*. Santa Barbara, CA: University of California, Santa Barbara. Retrieved from http://files.eric.ed.gov/fulltext/ED336901.pdf

Lau, A., Barnett, M., Stadnick, N., Saifan, D., Regan, J., Stirman, S. W., … Brookman-frazee, L. (2017). Therapist Report of Adaptations to Delivery of Evidence-Based Practices Within a

System-Driven Reform of Publicly Funded Children ' s Mental Health Services, *85*(7), 664–675.

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., Bishop, S. L., & Guthrie, W. (2012). *Autism diagnostic observation schedule–2nd edition (ADOS-2)*. Los Angeles, CA: Western Psychological Services.

Los Angeles County Department of Mental Health. (2010). Prevention and Early Intervention. Retrieved June 14, 2017, from http://dmh.lacounty.gov/wps/portal/dmh/!ut/p/b0/04_Sj9CPykssy0xPLMnMz0vMAfGjzOJdDQwM3P3dgo3cjd0cDTxdXYxD_AJMDC3NTPQLsh0VASrD6oM!/?1dmy&page=dept.lac.dmh.home.mhsa.mhsa_detail.hidden&urile=wcm%3Apath%3A/dmh+content/dmh+site/home/mental+health+services+act/

Mandell, D. S., Stahmer, A., 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

National Autism Center. (2009). *National standards report*. Randolph, MA: National Autism Center.

National Professional Development Center. (2010). Evidence-based practices for children and youth with ASD. Retrieved from http://autismpdc.fpg.unc.edu/

National Professional Development Center (NPDC). (2014). Autism focused intervention resources and modules. Retrieved from http://autismpdc.fpg.unc.edu/npdc-resources

National Research Council. (2001). Educating children with autism. *Washington DC: National*.

National Standards Report. (2009). *National standards report*. Randolph, MA: National Autism Center.

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. *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. (2010). Evidence-based practices in interventions for children and youth with autism spectrum disorders. *Education for Children …*.

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 and Developmental Disorders*, *45*(9), 2917–2927. https://doi.org/10.1007/s10803-015-2455-0

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.

Sanders, M. R., Markie-Dadds, C., & Turner, K. M. T. (2001). *Practitioner's manual for standard triple P*. Birsbaine, Austrailia: Families International Publishing. Retrieved from http://www.worldcat.org/title/practitioners-manual-for-standard-triple-p/oclc/48128950

Schoenwald, S. K., & Garland, A. F. (2013). A Review of Treatment Adherence Measurement Methods Sonja. *Psychological Assessment*, *25*(1), 146–156. https://doi.org/10.1037/a0029715.A

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

Schoenwald, S. K., Sheidow, A. J., Letourneau, E. J., & Liao, J. G. (2003). Transportability of Multisystemic Therapy : Evidence for Multilevel Influences, *5*(4).

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

Stadnick, N., Stahmer, A., & Brookman-Frazee, L. (2015). Preliminary effectiveness of Project ImPACT: A parent-mediated intervention for children with autism spectrum disorder delivered in a community program. J*ournal of Autism and Developmental Disorders*. *45*(7), 2092-2104. https://doi.org/10.1007/s10803-015-2376-y

Stahmer, A. C.**,** Brookman-Frazee, L., Rieth, S.R., Stoner, J.T., Feder, J.D., Searcy, K., & Wang, T. (2017). Parent perceptions of an adapted evidence-based practice for toddlers with autism in a community setting. *Autism: International Journal of Research and Practice, 21(2),* 217-230.

Stahmer, A., 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.* Guilford Press.

Stahmer, A., 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*. Retrieved from http://aasep.org/aasep-publications/journal-of-the-american-academy-of-special-education-professionals-jaasep/jaasep-winter-2016/a-pilot-examination-of-the-adapted-protocol-for-classroom-pivotal-response-teaching/index.html

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., Reed, S., Schreibman, L., Reisinger, E. M., & Mandell, D. S. (2013). Implementation challenges in translating pivotal response training into community settings . *Journal of Autism and Developmental Disorders*, *43*, 2970–2976.

https://doi.org/10.1007/s10803-013-1826-7

Wong, C., Odom, S. L., Hume, K. A., Cox, A. W., Fettig, A., Kucharczyk, S., … Schultz, T. R. (2015). Evidence-Based Practices for Children , Youth , and Young Adults with Autism Spectrum Disorder : A Comprehensive Review. *Journal of Autism and Developmental Disorders*, 1951–1966. https://doi.org/10.1007/s10803-014-2351-z

Young, H. E., Falco, R. A., & Hanita, M. (2016). Randomized, Controlled Trial of a Comprehensive Program for Young Students with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *46*(2), 544–560. https://doi.org/10.1007/s10803-015-2597-0

Table 1. *Participant Demographics*

Provider Characteristics	*N* = 23; *n* (%)	Child Characteristics	*N* = 19; *n* (%)
Gender		Gender	
Male	0 (0.0%)	Male	12 (63.2%)
Female	23 (100.0%)	Female	7 (36.8%)
Education		Mean Age in Months (*SD*)	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%)		
Asian	2 (8.7%)	Mean ADOS-2 Comparison Score (*SD; Range*)	7.50 (1.71; 4-10)
American Indian/Alaska Native	2 (8.7%)		
Not reported	6 (26.1%)	Receptive Language Age Equivalence Scores (in months)	
Ethnicity		Mean MSEL[1] (*SD*)	9.90 (4.93)
Hispanic/Latino	3 (13.0%)	Mean PLS-4[2] (*SD*)	23.75 (10.46)
Not Hispanic/Latino	14 (60.9%)		
Not reported	6 (26.1%)		
Setting			
Home	16 (69.6%)		
Classroom	7 (30.4%)		

Table 2. Coding Definitions and Reliability for Provider Behaviors

PRT Components	Definitions	Reliability for TBT (Cohen's Kappa)	Reliability for 5 Pt (ICCs)	Reliability for 3 Pt (ICCs)
Student Attention	Child is attending to the provider before the cue is provided either in proximity or orientation towards the provider.	.77	.39	-0.49
Clear Cues	Cue should be spoken in clear language or gestural expression.	.79	.23	.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.	.79	.86	.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.	.77	.78	.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.	.81	.41	.26
Turn Taking	Provider takes or facilitates turns while interacting with the child.	.86	.86	.03
Contingent Consequence		.80	.41	-.33
Direct Reinforcement	Provider uses contingent, tangible reinforcement for correct behaviors and attempts at correct responding, that is directly related to the teaching activity.	.81	.82	.94
Reinforcement of Attempts		.75	.86	.68
Reinforcement of Appropriate Behavior		.70	.95	.67

Table 3. *Coding criteria with descriptive anchors and % 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. Frequency data were then aggregated across each minute to facilitate comparison with other coding scales.	5 "Provider implements competently throughout the session" (100%)	3 "Provider implements competently most of the time, but misses some opportunities. Provider implements competently throughout the session.
	4 "Provider implements competently most of the time, but misses some opportunities." (80-99%)	
	3 "Provider implements competently half the time, but misses many opportunities" (50-79%)	2 "Provider implements competently occasionally, but misses many opportunities. Provider implements competently half the time, but misses many opportunities
	2 "Provider implements competently occasionally, but misses many opportunities." (30-49%)	
	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 (N/A) "Provider does not have the opportunity to implement during the session"	0 (N/A) "Provider does not have the opportunity to implement during the session"

Table 4. *Likert Scale ratings with the corresponding Trial-by-trial equivalent frequency percentages used to evaluate agreement.*

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

Table 5. *Percent of videos with agreement between the TBT and LS coding for individual PRT components*

	TBT to 5 point LS			**TBT to 3 point LS**			**5pt to 3pt**		
Component	Percent of exact agreement	Mastery Criteria Met agreement	KALPHA	Percent of exact agreement	Mastery Criteria Met agreement	KALPHA	Percent of exact agreement	Mastery Criteria Met agreement	KALPHA
Student Attention	100%	100%	1.0	72.2%	72.2%	.09	72.2%	72.2%	.09
Clear Cue	100%	100%	1.0	83.3%	83.3%	.29	83.3%	83.3%	.29
Developmental Appropriate	100%	100%	1.0	83.3%	83.3%	.30	83.3%	83.3%	.30
Shared Control	100%	100%	1.0	66.7%	66.7%	-.16	63.9%	63.9%	-.19
Maintenance/ Acquisition	94.4%	100%	.98	50%	63.9%	.46	44.4%	69.4%	.46
Turn taking	100%	100%	1.0	55.6%	63.9%	.49	27.8%	47.2%	.12
Contingent Consequences	100.0%	100%	1.0	72.2%	72.2%	-.13	55.6%	55.6%	-.15
Reinforcement of Appropriate Behavior	100%	100%	1.0	66.7%	75.0%	.59	38.9%	52.8%	.15
Direct Reinforcement	100%	100%	1.0	72.2%	75.0%	.08	55.6%	66.7%	.56
Reinforcement of Attempts	100%	100%	1.0	44.4%	52.8%	.24	61.1%	63.9%	<-.01