


# Measuring Fidelity in Research Studies: A Field Guide to Developing a Comprehensive Fidelity Measurement System

Megan Feely<sup>1</sup>  · Kristen D. Seay<sup>2</sup> · Paul Lanier<sup>3</sup> · Wendy Auslander<sup>4</sup> · Patricia L. Kohl<sup>4</sup>

Published online: 16 August 2017  
© Springer Science+Business Media, LLC 2017

**Abstract** An accurate assessment of fidelity, combined with a high degree of fidelity to the intervention, is critical to the reliability, validity, replicability, and scale-up of the results of an intervention research study. However, extant measures of fidelity are infrequently applicable to the program or intervention being studied, and the literature lacks guidance on the specific process of developing a system to measure fidelity in a manualized intervention. This article describes a five-step process to define the scope, identify components, develop tools, monitor fidelity, and analyze outcomes to develop a comprehensive fidelity measurement system for an intervention. The process describes the components, measures and key decisions that form a comprehensive fidelity measurement system. In addition, the process is illustrated by a case study of the development of a fidelity measurement system for a research study testing Pathways Triple P, a behavioral parent-training program, with a population of child welfare-involved families. Pathways Triple P is a common, manualized intervention and the process described in this article can be generalized to other manualized interventions. The implications and requirements for accurately assessing and monitoring fidelity in research studies and practice are discussed.

**Keywords** Interventions · Fidelity · Measure development · Measures · Triple P · Pathways Triple P · Fidelity adherence · Treatment adherence · Implementation

## Introduction

Measuring and monitoring the degree of fidelity in research is critical to establishing the evidence base of interventions and determining the circumstances under which an intervention is effective. Interventions need to be delivered with a high degree of fidelity to the model to ensure that the results of the intervention reflect a true test of the program. There is general agreement and substantial literature on two points related to the assessment of fidelity:

1. Fidelity should be measured in intervention research studies. (Dane & Schneider, 1998; Moncher & Prinz, 1991; Proctor et al., 2011; Schoenwald, 2011) and
2. An effective fidelity measurement tool should incorporate key components (Carroll et al., 2007; Cross & West, 2011; Proctor et al., 2011).

The NIH Behavior Change Consortium provides an excellent overview of how to incorporate fidelity measurement into the design and delivery of a research study (Bellg et al., 2004). A recent publication outlining the state of assessing fidelity in interventions within the child welfare system reveals that the field recognizes the importance of fidelity but fidelity is integrated inconsistently into studies and usual care (Seay et al., 2015). However, the literature does not provide sufficient detail to guide practitioners in the development of a comprehensive system for monitoring the degree of fidelity in a research study.

---

✉ Megan Feely  
megan.feely@uconn.edu

<sup>1</sup> University of Connecticut School of Social Work, Hartford, CT, USA

<sup>2</sup> College of Social Work, University of South Carolina, Columbia, SC, USA

<sup>3</sup> University of North Carolina School of Social Work, Chapel Hill, NC, USA

<sup>4</sup> Brown School at Washington University in St. Louis, St. Louis, MO, USA

Investigators need accurate and useable measures to assess the degree of fidelity in research studies (Schoenwald & Garland, 2013). Intervention manuals are generally designed to facilitate the training process and to be used as a resource for practitioners throughout the intervention. They are not designed to be scorecards of how closely practitioners adhere to the model. Therefore, they may or may not contain information and measures that are sufficient to assess the degree of fidelity. Researchers may need to develop intervention-specific fidelity measures. Currently, there are no clear guidelines on how to translate a manualized intervention into an accurate fidelity measure. This paper fills this gap by providing a step-by-step process for developing a fidelity measurement system. The five-step process that was developed, *The Field Guide to Fidelity* (hereafter referred to as the *Field Guide*), is a flexible process tool that can be used to create fidelity monitoring systems for manualized interventions. To enhance the usefulness of the *Field Guide*, this paper details key decision points, measurement tools that may need to be developed, ways to score and analyze the fidelity ratings, possible options for how to structure the measurement tools and scoring, and some of the implications of different choices. These details are generally incomplete or absent in other descriptions of fidelity measures but they are critical to the development of a comprehensive system. The final fidelity measurement system may include several tools to measure different components of fidelity. The application of the guide is illustrated through a case study of one research study testing a manualized parenting intervention, specifically the Pathways Triple P program (PTP) (Sanders et al., 2003b; Sanders et al., 2004; Turner et al., 2002). The fidelity system developed for PTP can also serve as a tool for researchers studying other variants of Triple P.

### Pathways Triple P Case Study

The fidelity measurement project was a component of a larger randomized control trial testing the effectiveness of the PTP behavioral intervention with families who had been referred to the child welfare system following an allegation of physical abuse or neglect and whose child remained in-home after an investigation or assessment. PTP is part of a continuum of Triple P parent support and training programs that proscribe multiple levels of intervention varying in intensity (Sanders et al., 2003a). PTP was developed for parents at risk of maltreating their children and combines parenting skill training with techniques designed to reduce negative parenting beliefs, parental anger and stress. The purposes of measuring fidelity in this study were to ensure that the intervention was delivered as designed, and to determine whether the level of fidelity impacted treatment outcomes. The case study follows the development and

utilization of the fidelity monitoring system through all five steps of the *Field Guide*.

### Understanding the Importance of Fidelity

Fidelity is defined as “the degree to which teachers and other program providers implement programs *as intended by the program developer* (emphasis in original)” (Dusenbury, Brannigan, Falco & Hansen, 2003, p. 240). Assessing fidelity against a model increases the reliability and validity of the results of a behavioral intervention, because it ensures that all participants are receiving the same intervention (Schoenwald et al., 2011). If all participants receive the intervention components in the prescribed manner, then the outcomes can be attributed to the intervention. Without this assurance, the connection between the results and the intervention is unclear (Dusenbury et al., 2003; Moncher & Prinz, 1991; Miller & Rollnick, 2014). While this is the most basic definition and purpose of fidelity, a more nuanced and complete definition is needed to develop a fidelity monitoring system.

### Frameworks and Theories of Fidelity

The various frameworks and theories of fidelity tend to advocate one of two different approaches to fidelity measurement (Cross & West, 2011; Moncher & Prinz, 1991). One category focuses solely on the interaction between the individual client and the provider in a manualized intervention. The other takes a broader view, and incorporates organizational variables that may impact fidelity. However, many of the underlying concepts between these two frameworks are similar.

To assess individual-level interaction, Proctor and colleagues identified five dimensions of fidelity: adherence, exposure to the intervention, quality of delivery, component differentiation, and participant responsiveness or involvement (Proctor et al., 2011). Adherence is whether the components of the intervention are delivered as intended. Exposure, sometimes referred to as dose, measures how much of the intervention was delivered. Quality of delivery evaluates provider skill and competence according to the manner of delivery specified in the intervention manual or training.

Component differentiation is important when the intervention specifies a particular mode of delivery and excludes the use of other practices not specifically taught as part of the training. For example, if the principles of cognitive behavioral therapy (CBT) are not specified in the intervention, then practitioners should avoid using CBT with clients. Including CBT in the intervention could confound the study results because it would be unclear whether the intervention content being

tested was successful, whether the CBT techniques were successful, or whether it was the intervention content with CBT techniques that was successful. Finally, participant involvement assesses whether the participants are actively engaged in the learning process.

Other frameworks for fidelity measurement incorporate agency or organizational level factors (Cross & West, 2011). For example organizational variables could include caseloads or the level of training, education, and experience of individuals who will be delivering the intervention. Failing to maintain those conditions (i.e., fidelity to treatment protocols at the organizational level) may result in less successful treatment outcomes when the intervention is tested in an effectiveness trial. Therefore, the *Field Guide* example here includes fidelity processes that include individual and organizational-level factors.

### Field Guide to Fidelity Measurement

The *Field Guide* describes a sequential five-step process displayed in Fig. 1. The steps are:

1. defining the purpose and scope of the fidelity assessment used for evaluation of the intervention;
2. identifying the essential components of the fidelity monitoring system;
3. developing the fidelity tool;
4. monitoring fidelity during the study; and
5. using the fidelity ratings in analyses.

Each step includes key decision points and examples from the PTP case study. Decisions in the earlier steps inform the later steps, so the *Field Guide* is best used in the order presented.

#### *Field Guide Step 1: Determine Purpose and Scope*

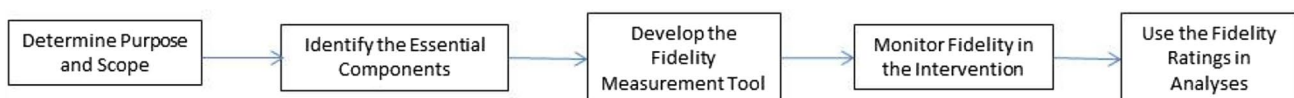
The first step in designing a fidelity monitoring system is to determine its purpose (Fig. 1). This includes considering what the information will be used for, whether the fidelity assessment will focus only on the individual level or also include the organizational variables, and how much information can and should be gathered. Time and budget constraints must be considered because the development and measurement of fidelity can take up a significant portion of a project's time and budget. To minimize the burden of fidelity data collection, it is

important to establish the scope of the fidelity component in the early stages of the study design to ensure that the key information is gathered but extraneous information is excluded.

Measuring fidelity has multiple uses within a research study. First, it can be used to confirm what is being delivered to the clients. Strong adherence to fidelity standards makes it possible to avoid a “Type III” error, which occurs when the results show no significant effects but the intervention was not delivered with consistently high quality (Dobson & Cook, 1980). Without strong fidelity, the results are inconclusive and it is unclear if the intervention was ineffective or if it would have been effective if properly delivered (Dobson & Cook, 1980). Second, a fidelity tool can guide supervision and help deliver the intervention consistently over time (i.e., avoid drift). Third, fidelity ratings assist in the overall analysis of the intervention. Such ratings permit consideration of whether results were affected by the degree of fidelity to the model.

The scope of a fidelity assessment system should match the needs of the study and provide key information for any planned future research projects. In an efficacy trial designed to establish if an intervention is successful under highly controlled conditions, fidelity monitoring may focus more on individual-level variables, specifically whether the practitioner delivered the intervention correctly. However, other characteristics of the implementation process that may be necessary to replicate the results should also be recorded and reported in publications, such as the frequency and purpose of supervision, the level of training of the practitioners, and caseload size. Furthermore, variability in these implementation characteristics can influence outcomes. Hence, in multi-site intervention studies they should be assessed and used as site characteristics in analyses that account for clustering by site. A clearly defined purpose and scope of fidelity measurement should guide decisions in the rest of the fidelity rating system.

*Pathways Triple P Case Study* The purpose of monitoring the degree of fidelity in the PTP study was to determine the extent to which the practitioners were delivering the intervention as intended and to use the measure of adherence as a moderator of treatment outcomes in the analyses. The scope of the fidelity monitoring system was written into the grant, which was guided by existing literature on components of fidelity measurement. The research team identified additional tools that needed to be developed.



**Fig. 1** Field guide to fidelity model

Consistent with the NIH suggestions (Bellg et al., 2004), the research plan included the key components to support fidelity measurement. Specifically, the plan included organizational-level variables, such as the training, certification and supervision of the practitioners, the qualifications of the supervising practitioners, and the caseload size (Bellg et al., 2004). There was also an existing self-report measure of participant engagement that was administered at three time-points during the intervention. Therefore, to have a comprehensive system the only tools needed to complete a comprehensive fidelity measurement system were measures to assess the process and content of the practitioner/client interaction. Assessing component differentiation was intentionally left out of the plan. The Triple P intervention system allows practitioners to use therapeutic skills to develop a collaborative relationship with the parents (Sanders et al., 2001). Therefore, the team decided not to specify these additional techniques or components used by the practitioners.

### *Field Guide Step 2: Essential Components*

**Content and Process** The second step in the *Field Guide* is to identify the essential components to measure. In the individual level interactions between practitioner and client, fidelity measurement may focus primarily on content, i.e. the informational parts of the intervention (the “what”), or it may also include the process through which the intervention is delivered (the “how”).

Adherence to the content is often considered the primary objective in measuring fidelity (Carroll et al., 2007; Cross & West, 2011). The content of the intervention “may be seen as its ‘active ingredients’ such as the drug, treatment, skills or knowledge that the intervention seeks to deliver to its recipients” (Carroll et al., 2007, p. 4). Depending on the intervention, the content may be presented as specific steps, as key information to be delivered, or as more general concepts to convey to the participant. The initial list of core concepts or specific steps should be identified using the manual, other training materials, and any existing literature on measuring fidelity in the intervention. Enough such items should be identified so that the assessment process will furnish sufficient detail to make an accurate determination of the degree to which the intervention remained faithful to the model. Using too few components or components that are defined too broadly would limit the usefulness of the fidelity measurement tool, because scoring would not accurately reflect adherence to the model. However, including too many items may be unnecessary and burdensome to the raters. The list of possible items may need to be edited to exclude superfluous or less important steps. If component differentiation is going to be measured, a means must be developed to account for information or techniques that are beyond the scope of the intervention. A useful measure of content fidelity should

capture the full range of content implementation so that it may be used in analyses as a quantitative measure of the proportion of the intervention that is correctly delivered.

Many different terms are used to describe the process or manner in which the intervention is delivered including: process fidelity (Dumas, Lynch, Laughlin, Smith, & Prinz, 2001); competence (Madson & Campbell, 2006); clinical processes or competent execution (Forgatch, Patterson, & DeGarmo, 2006); quality of delivery (Carroll et al., 2007; Schoenwald, 2011); therapeutic alliance (Beidas, Benjamin, Puleo, Edmunds, & Kendall, 2010); interventionist competence (Cross & West, 2010); and consultation skills (Mazzuchelli & Sanders, 2010). All of these terms describe whether “the intervention is delivered in a way appropriate to achieving what was intended” (Carroll et al., 2007, p. 6). The process is important to the proper implementation of an intervention (Hamilton, Kendall, Gosch, Furr, & Sood, 2008). Following the process described in the intervention should create an environment that allows the participant to learn the information (i.e., content) that is being delivered (Forgatch et al., 2006). Higher process ratings may predict positive treatment outcomes (Forgatch et al., 2006). In controlled trials, treatments are usually delivered by trained clinicians with extensive experience. In effectiveness studies of implementation in usual care, the training and experience of practitioners tends to be more varied. For effectiveness studies to be faithful to the model, clinicians need to adhere to the process as outlined by the treatment developers and demonstrated in clinical trials.

Identifying the essential components for content and process are important steps in developing a fidelity measurement tool, and will likely be an iterative process that involves people trained in the intervention, literature on the intervention, and the treatment manual. In some cases, and where available, consultation with the model developer may be advised.

**Participant Responsiveness** Participant responsiveness assesses how the participant is engaging in the intervention (Dane & Schneider, 1998). Topics to measure may include participants’ enthusiasm for the intervention, comprehension of the information, and application of the skills (Dane & Schneider, 1998; Gearing et al., 2011). These can be assessed through self-report by the participant, or by the clinician, or by an objective rater listening or viewing a recording. Specific topics of responsiveness may include the participant’s openness to the intervention, his or her level of participation in an individual session, the degree to which the participant has integrated the teachings into his or her life, and how much effort the participant has exerted to work for change, as represented by concrete tasks such as completing homework.

*Pathways Triple P Case Study* The essential components of the intervention were ascertained via training and accreditation in the intervention model, review of the published material, and expert consultation. The fidelity assessment development team attended the week-long PTP training. The training provided important grounding in the philosophy of the PTP system, familiarized the fidelity development team with the materials the practitioners used, and taught them the key process components of PTP. The fidelity team was led by a doctoral student trained and accredited in PTP. In addition to training on the specific content, the training focused extensively on the interaction between the parent and practitioner. Through teaching and practice sessions, the practitioners developed the skills necessary to lead parents through the intervention.

**Content:** The PTP model has content specific to each session and the material builds from week to week. To achieve the full benefit from the intervention, it is important for the parent to be taught all of the parenting skills integral to the program and have a chance to practice the skills under the practitioner's guidance and supervision. Therefore, the essential content was defined as all of the individual steps outlined in the manual. This was an extensive list of items, but all of the items were deemed essential in aiding the parent to develop the new skills.

**Process:** The PTP training specifies ways for the practitioner to interact with the parent and deliver the intervention that supports the development of the parent's self-regulatory framework. This framework suggests that parents: (1) decide which of their own behaviors and which of their child's behaviors they would like to change, (2) set personal goals, (3) choose which parenting techniques they would like to implement, and (4) be encouraged to self-evaluate their progress toward their goals and success with the chosen techniques (Sanders, Turner, & Markie-Dadds, 2002). One of the main goals of PTP is to teach a parent to find her own solutions for her child. By doing so, the parent develops the confidence and capacity to solve various parenting challenges. To be faithful to the PTP program, practitioners need to promote the development of these skills in the parent.

The PTP manuals and training course include extensive discussion of the essential process components of PTP. The self-regulatory framework that guides the delivery of PTP and that served as the foundation for our process fidelity measurement tool is laid out clearly in the literature (Sanders et al., 2002), as well as in the treatment manuals (Sanders & Pidgeon, 2005; Sanders et al., 2001). Yet, distilling lengthy descriptions and examples from the manuals into measurable categories that should occur in every session was a challenging process. The process categories that were consistently emphasized in training and in the manual and supported by a PTP expert were identified by the fidelity measurement development team for inclusion in the fidelity measurement

tool. Specific steps for developing the list of process items are discussed in the next section.

**Participant Responsiveness:** Participant responsiveness and engagement was assessed through a nine-item scale developed by the study's clinical team. The practitioners filled out the responsiveness measure after every session.

### *Field Guide Step 3: Developing Fidelity Measurement Tools*

The third step in the *Field Guide* is to develop the fidelity measurement tools. These tools measure the degree of adherence to the manual (Moncher & Prinz, 1991). A number of key measurement-related criteria need to be considered when developing a new measure or modifying an existing measure: (1) the organization of the fidelity measurement tool; (2) items to be included on the list; (3) phrasing of the items; and (4) response choices.

*Design of a Fidelity Tool* The structure of the intervention and delivery should inform the design of the tool. For example, the intervention may be structured as a set of skills the participant should master before moving on to new skills, or it could be structured as a set number of sessions to deliver, regardless of mastery. The intervention may be delivered in groups or to an individual, in the client's home or in the professional's office. Each of these various modes of delivery and structures may require a slightly different fidelity measurement tool. For example, if the structure requires a participant to master a technique or concept before moving to the next section, then the fidelity measurement tool should be designed around the skills or techniques that the participant has to learn. However, if each session mandates the delivery of specific material, then designing the tool around the sessions is more logical.

The skill level of practitioners also needs to be considered (Cross & West, 2011). If practitioner skill and on-going supervision vary, then it is likely that some practitioners will be very well trained and supported and others will be less so. Hence, a more detailed measure might be needed to capture this variation. A more detailed measure would divide key aspects of process into more nuanced components to capture a more accurate picture of the degree of fidelity across the full range of implementation.

Developing a fidelity measure for interventions that are structured around skill accomplishment may require a more complex and flexible fidelity tool because it needs to allow different skills to be introduced at different times. The Fidelity of Implementation Rating System (FIMP) for the Oregon model of Parent Management Training (PMTO) is an example of a fidelity measurement and monitoring tool for an intervention structured around skill acquisition (Forgatch et al., 2006). PMTO is delivered by highly trained clinicians

who are closely supervised throughout the delivery; and fidelity is continually monitored by similarly skilled clinicians who have also undergone extensive training and supervision for fidelity monitoring (Forgatch et al., 2006). However, this may not be a feasible model for all interventions.

*Deciding What to Include* Before developing a new measure, it is important to assess the quality and utility of any existing tools for measurement. There are many preexisting measurement tools, and one may have been developed in conjunction with the chosen intervention, or may be modifiable from a similar intervention (Schoenwald & Garland, 2013). Treatment manuals often have a self-assessment or reminder checklist for the practitioner. Checklists in the manual, practitioner self-assessments or formal fidelity measures can provide a helpful foundation for a more detailed measure to accurately assess fidelity in a research study. If the intervention lacks an existing fidelity measurement tool, then one can be developed through an iterative process. Several people trained in the intervention should develop lists of components that should be measured. The lists should be compared, guided by the manual, by received training, and by expert opinion. The final list should be narrowed down to the important steps. The number of items to be measured should generally reflect the specificity of the manual and the training. There should be more listed items for interventions that specify many details of the intervention and delivery, and fewer for interventions with less detailed directives. A list with too few items, however, makes it difficult to distinguish a thorough intervention delivery from an incomplete one. A more detailed measure will more accurately measure fidelity, which will in turn be more useful for assessing the intervention and in analyses.

*Phrasing of Items* Each item should measure only one topic or type of process. When questions clearly measure one item (e.g., whether the practitioner defined time-out) it is clear what a negative score on the fidelity measure means. If the items are double-barreled, e.g., “the practitioner covered the topics of timeout and positive reinforcement”, it is unclear what a negative answer means. There are three possible combinations of information that would result in a negative answer to such a question. The combinations are (1) timeout was covered but not positive reinforcement; (2) positive reinforcement but not timeout; (3) neither was covered. If there are many multiple choice questions such as this, the participant could have received as much as half the information in the intervention or as little as none of it. Similarly, intensity and accuracy should be assessed in separate items rather than in a compound question. For example, a negative answer to a question such as “the practitioner explained timeout in a clear and patient manner” is confusing. The practitioner could have explained timeout

in an impatient and/or unclear way, not explained at all but been patient, or only vaguely referred to timeout and been impatient and unclear. Some measures may try to account for different aspects of the answer by having very detailed process categories that specify these combinations (e.g., one answer option is that the practitioner explained clearly but was impatient, another option that practitioner was unclear but was very patient). However, these measures may be less useful in analyses because they do not produce a continuous variable.

*Selecting the Response Options* The most common types of response options are a binary yes/no answer, a Likert scale with three to seven answer options, or a continuous scale that captures the whole range of the answer possibilities. The yes/no, or completed/did not complete answer option may be useful for content items and practitioners need to deliver all content to receive credit.

If it is only important to deliver limited information about general topics, then a Likert scale may be useful. The answer options could be on a none-to-all range to reflect the amount of information that was delivered for that specific topic. A Likert scale could also be useful in a process measure to capture the practitioner’s adherence to a process item. The third type of response scale is a 10 or 11 point scale often used in conjunction with a phrase-completion question. This type of response option is less common, but it more accurately captures the underlying construct and produces more variation than a Likert scale (Hodge & Gillespie, 2007). An example of this concept is a scale of how often the practitioner demonstrated a process item. A phrase-completion scale asks about the frequency of desired (or undesired) behaviors on a scale that goes from 0 to 10 or 0 to 100 and only has anchoring statements at the ends. For example, a process-related phrase-completion statement might be, “The practitioner was patient with the parent” and the answer options could range from “none of the time” to “all of the time” on a scale from 0 to 10. The rater would select the number on the scale that represents the percent of the time that the practitioner was patient.

*Pilot-testing the Measure* After developing the best measure, with clearly defined single-topic items and appropriate response options, the measure should be pilot-tested for clarity and consistency. To pilot-test the tool, at least two people who are familiar with or trained in the intervention need to rate several sessions. The testers should have a specific goal of percent agreement (i.e., the percent of the total fidelity tool items that matched between the fidelity raters), such as at least 90%. During the pilot phase it is simpler to calculate percent agreement rather than a kappa or inter-class correlation. At this stage, all discrepancies in rating should be discussed. If some raters are interpreting the tool

differently, then the problematic items should be clarified, reworded, split into two or more separate items, discarded or otherwise changed. The revised tool should then be piloted again to test the new items. This may require several iterations to achieve a final measure, and user feedback may be useful.

*Pathways Triple P Case Study*

PTP is structured for specific content to be delivered in each session. The manual contained checklists for each session (Sanders et al., 2001; Sanders & Pidgeon, 2005) that were used as a starting point for the measure of content fidelity. These checklists listed all the steps that a practitioner should follow (e.g., review the agenda for the session) and the information that needed to be conveyed (e.g., introduce quiet time). The final fidelity measures for the project consisted of 14 content checklists specific to each of the 14 sessions, and one process checklist that was for every session. Each content item was rated as either present or absent (a yes/no binary response) and each process item was rated on an 11-point phrase-completion scale.

*Content Items* As noted above, the starting point for each session’s content was the checklist provided in the manual. After a brief pilot-testing of the original checklists, the project team determined that greater specificity was needed to obtain reliable measures of fidelity. Table 1 presents the decision points and the team’s choices to improve the usa-

bility of the tool. Figure 2 provides an example of a content item from the checklist.

*Process Items* The list needed to capture the core process components of PTP to produce scores that were consistent with an expert assessment of a good session, and to have high inter-rater reliability. The items were identified from the process concepts in the PTP manual and training and were expanded for clarity. There were no “penalties” for incorporating other techniques to develop a collaborative relationship with the parent (Sanders et al., 2001).

The items on the final process checklist are:

1. Reflective listening—this includes verbal validation of the parent.
2. Using a non-judgmental tone and non-judgmental questions when talking with the parent.
3. Allowing the parent to identify her own solutions.
4. Asking open-ended questions.
5. Not interrupting the parent.
6. Encouraging the parent throughout the session.
7. Managing time in the session.
8. Tailoring the content of the session to the parent’s needs and/or abilities.

To facilitate the ease of rating process fidelity, the final process measure clarified each item by providing additional detail and examples of what the practitioner might say to address the item (Fig. 3). Only the first seven items were

**Table 1** Decisions on the PTP Study: content and process checklists

Decision points	PTP team decisions
Design of the fidelity tool	Each session had a customized and detailed list of content items and the process items were consistent across all sessions. This mirrored the design of the PTP intervention
Phrasing of items—single concept	Items that asked about two tasks (e.g., ask about homework and review agenda) were split into two items so that it was clear when the practitioner had completed the task
Phrasing of items—clarity	The checklists were designed to be used by trained raters without constant reference to the manuals so jargon was rephrased in common language and all items that referred to the manual (such as “Complete Activity 2”) were expanded to include the key information from the manual
Selecting the response options	For the content, a binary yes/no option for ‘completed’ and ‘did not complete’ was used. A 0–10 phrase completion scale with the statements at the ends being, “None of the appropriate times” and “All of the appropriate times”
Pilot testing the measure	The measure, particularly the process section, went through several rounds of testing until the raters reached about 90% agreement and were consistently within 1 point on the phrase completion scale. Examples were added for any content items that continued to be confusing and for all process items

1. Agenda			Time*	Item number
Provide an overview of what will be covered in the session (Practitioner may go over session goals or other parts of the agenda)	Yes	No		1

**Fig. 2** Example of one item from the process checklist

included on the first version of the process checklist; however, the raters were not able to consistently achieve agreement. The eighth item, tailoring the content to the parent’s needs and/or abilities, was added later to distinguish poorly managed sessions from sessions where the practitioner made adjustments in the delivery to accommodate challenges with the parent or environment. For instance, a few parents had cognitive impairments which impacted how the practitioner presented the content, such as providing additional examples, and adjusting the rate at which practitioners presented the intervention, usually taking more time to cover each concept. In other sessions, children or other adults in the home were very disruptive, so the practitioner adjusted the delivery or content to complete the session.

These adjustments were within the scope of the flexible delivery that Triple P allows and endorses (Mazzucchelli & Sanders, 2010); however, when rating the process of these sessions, each rater attempted to quantify the practitioner’s adjustment in different categories. This resulted in inconsistent inter-rater reliability. For example, in a situation with a parent with some cognitive problems, one rater might subtract points on *managing time in the session* but give credit to the practitioner for *encouraging*. For the same situation, another rater would not penalize the practitioner on the *managing time* item but might credit the practitioner in the *not interrupting* item. After extensive discussion of the discrepancies, the team added the eighth item, which improved the rater’s agreement.

**Response Options** The PTP team decided to use a yes/no response for the content items. The item had to be entirely completed to receive a “yes” rating; no partial credit was awarded. The scoring was simple; the total number of “yes”

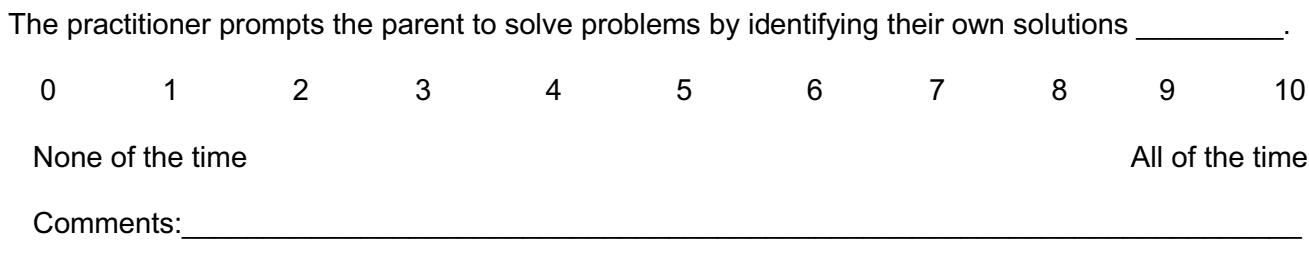
answers divided by the total number of items for the session was the percent of the content delivered. This was the first system that was tried and it was easy to use and produced consistently good inter-rater agreement.

Developing the response scale for the process items was more complicated. The first iteration of the process scale was a three-category Likert scale of “none of the time”, “some of the time” and “all of the time.” This scale did not produce any variability in the responses, because the practitioners never “asked open ended questions” at every single opportunity, but they always asked them some of the time. To enhance the accuracy of the measure, a phrase-completion scale was selected (Hodge & Gillespie, 2007). The scale’s range was 0–10 with the anchor at 0 phrased as “None of the appropriate times” and the anchor at 10 as “All of the appropriate times”. An example is provided in Fig. 3. The rater selects the appropriate number on the scale to reflect how often the process measure was utilized, relative to how often the process measure could have been utilized. This scale produced variability in the process measure and consistent scores in the piloting phase. Additionally, it had better face validity, meaning that sessions that sounded like “good sessions” had higher scores than sessions that had some marked weaknesses.

*Field Guide Step 4: Monitoring*

The fourth step in the *Field Guide* is to determine how fidelity will be monitored and rated. This step includes the technical aspects of recording the sessions, identifying which sessions to rate, and the personnel who may be involved in the rating process.

Triple P encourages parents to develop problem-solving skills. Practitioners can help parents develop these skills by asking them about other ways they might handle situations, or prompting them to think about ways they have successfully dealt with a problem in the past. If the practitioner is overly directive, then the “penalty” is assessed in this category.



**Fig. 3** Example from the PTP process measure



*Methods for Rating* There are four main methods that allow for the fidelity of a session to be rated: (1) direct observation in-person; (2) self-report of the practitioner; (3) audio recording; or (4) video recording. All the options except self-report involve an independent rater of fidelity. Direct observation is the most expensive and complicated, because it requires the second person to be present when the intervention is delivered. However, it has been used for supervision and quality control by SafeCare (Aarons, Sommerfeld, Hecht, Silovsky, & Chaffin, 2009). Although self-report is the least expensive and easiest option, it is also the most susceptible to bias because practitioners assess themselves. With audio and video recording, some or, preferably, all of the sessions will be recorded. An independent rater can listen to or watch the session to rate the degree of fidelity based on what he saw and/or heard in the recording of the session. For most interventions, audio or video recording provides a balance between complexity of scheduling, cost, and accuracy. The recordings can be stored electronically, reducing the need for physical storage space, and can be reviewed in a lab or office at a convenient time rather than scheduling an additional person at the session. Recording all sessions has two benefits over recording only some sessions or other methods. If all sessions are recorded, the practitioner is unaware of which sessions will be rated, so it is a more accurate sample of how the intervention was delivered. Second, even if initially it is unfeasible to rate all the sessions, the data are there and can be rated later to provide fidelity ratings for every session.

*Personnel Selection and Training* Recruiting and training raters is an important step in the process. Ideally, raters would be detail-oriented and willing to commit to completing all the ratings so that new raters do not have to be trained and the raters are consistent across sessions. Fidelity raters need to be familiar with the intervention and should receive thorough training in using the fidelity measures, with a focus on practical examples. Ideally, the raters should be trained in the intervention; but alternatively, a shorter training focused on the fidelity rating could be used. Ratings should be conducted independent of input from the practitioners, to preserve the objectivity of ratings.

Rating can occur concurrently with the intervention or after all sessions are completed. There are advantages to each approach. If the rating is concurrent with the intervention delivery, the fidelity ratings can be used to assess whether the delivery is consistent throughout the intervention, or if there is some drift over time (Moncher & Prinz, 1991). This could supplement or replace clinical supervision. However, waiting until all sessions are completed is more efficient for rater training, and more likely to produce

consistent scores. If this is the rating method chosen, then all clients must have completed the intervention so that the pool of possible sessions to rate has been identified.

A key decision is whether consistent ratings or testing the reliability of the fidelity measure is more important. If consistent ratings are more important, then the raters will want to compare scores frequently to make small adjustments. If it is important to test the reliability of the fidelity measure and whether it can be used to achieve consistent scores by different people, then fewer comparisons should be made to preserve the validity of the kappa value. Further discussion of consistency and kappa calculations is in a later section.

*Randomization* The sessions to be rated should be randomly selected unless all sessions will be rated. The decisions related to randomization are: (1) the percentage or number of sessions to be rated; (2) how many will be rated by two raters; (3) the randomization method that accounts for participants failing to finish all the sessions; (4) any clustering or stratification that is inherent in the study design. The number of sessions that are rated will depend on the budget and the type of intervention. Each session or type of session should be represented approximately equally so that the final fidelity score accurately reflects fidelity for the entire intervention. Likewise, if there are multiple practitioners, each practitioner should be represented equally in the sessions that are to be rated.

One key consideration in the randomization process is that for interventions delivered to one person at a time sessions that occur earlier in the sequence of the intervention are more likely to occur because participants may drop out before reaching the later sessions. This has two implications for randomization. First, if sessions are randomized and the sessions to be rated are selected in the beginning of the project, the randomization process will select some sessions that may not occur. All intervention studies have some participants drop-out, consequently all participants will not complete all of the sessions in the intervention. Second, for an individual intervention with specific content for each session, selecting a set number of participants from each session to rate (e.g., ten) ensures that the later sessions, after some participants have dropped out, have the same contribution to the fidelity score as the earlier sessions. This is less of a concern for group interventions because as long as one person is still attending the group, fidelity can be rated. Additionally, the selected sessions should be approximately representative of the study population. Therefore, depending on the study design, stratifying the randomization process by practitioner, site or client characteristics may be necessary to ensure that the final sample is representative of the sample population.

### Pathways Triple P Case Study

The key decision points and the project decisions are presented in Table 2. Every session was recorded, so every client interaction was available for review in supervision and to be randomized. Additionally, the practitioners did not know which sessions would be rated for fidelity. Because of attrition, the pool of possible sessions for randomization was not identifiable until all the clients had completed the sessions. The same number of episodes, eight, of each session were selected rather than a percent of the total. Rating a percent of the total number of sessions (e.g. rating 10% of all sessions) would have over-represented earlier sessions.

While there were four practitioners that delivered the intervention, clients were not distributed equally among practitioners. Because of staffing patterns, one practitioner (i.e., Practitioner A) was assigned to about half of the participants and the other three practitioners were assigned to the other half (i.e., Practitioners B, C, D). A block randomization within each session by practitioners was utilized to provide a balance of episodes between the practitioners. This was done using the random number generator in Excel. All the sessions were assigned a random number, the numbers were ordered, and the first four episodes were selected from each of the two blocks. Specifically, for each session four participants were chosen from Practitioner A and four from Practitioners B, C, and D, for a total of eight participants.

All fidelity raters underwent 15 h of training. The training included an overview of the intervention, listening to examples from different sessions, and practice in scoring sessions to reach inter-rater agreement. Sections of recorded sessions were used to provide real examples of strong fidelity to the model, and of interactions that were not as faithful to PTP. The same sessions were used to develop the fidelity measure and during the training to preserve as many sessions as possible for the randomization. The fidelity trainer also reviewed the content of each session and specific details from the manual with

the raters. Four raters were trained and assigned to rating pairs. Each pair had one trained professional with extensive clinical experience. These pairs were constant throughout the rating.

All the raters scored the same sessions until the scoring was above 90% agreement with the original scoring by the fidelity tool development team. Then, the raters scored sessions with their pair until the pair had 90% agreement. Finally, they scored the sessions selected for rating through the randomization process. The research team decided it was more important to have consistent ratings than to test the inter-rater reliability of the measure. Therefore, there were frequent comparisons between raters. When a new session was started, both raters would score a session, then compare their scoring item by item. Then they would score a second session, making any minute adjustments in their judgments to try to reach closer agreement, and then would compare answers again. If they achieved at least 85% agreement on the second session, then one rater would score the other six sessions. If they were still below the 85% mark, then the protocol was for both raters to score the next (i.e., third) session. The second, and more experienced rater, scored one of the final six sessions, but the first rater was blinded to which one was being scored until all episodes for that session were scored. This system aimed to increase agreement while still retaining validity for inter-rater agreement and minimizing the number of sessions with two raters. A fidelity measure development team member was available for consultation throughout to address any points of confusion.

### Field Guide Step 5: Use of Fidelity Ratings in the Outcome Evaluations

The final step in the *Field Guide* is how to use fidelity ratings in the analysis of the outcome data. In addition to confirming that the intervention was delivered as intended, using fidelity data in the analysis can be used to understand the relationship between the quality of the delivery and the effectiveness of the intervention. For example, fidelity ratings can be used as a moderator in the relationship between the intervention

**Table 2** Decisions on the PTP Study: rating the sessions

Decision points	PTP decision
Method for rating	Sessions will be rated using the digital audio recording of the session
Recorded sessions	All sessions were recorded
Randomization process	Eight episodes of each session were randomly selected at the end of the study from the list of completed sessions. The randomization was stratified by practitioner
Raters	There were two teams of raters, each with a Masters student and a PhD or PhD student. One team rated all even numbered sessions, the other odd numbered (e.g., all Session 2s were rated by the pair assigned to even numbered sessions)
Training for raters	The raters were trained in a two-day training that included listening and practice rating sessions. Each team had to achieve above 85% agreement on training sessions before beginning rating

and the intervention outcomes. Complete and nuanced fidelity data allows for a more comprehensive analysis of the intervention's effectiveness than would be feasible with more cursory data or simply dichotomous categories of treatment and no treatment (Moncher & Prinz, 1991). Moreover, in cases where the data indicate that the intervention is not effective, fidelity ratings are extremely important in understanding some of the reasons for a lack of significant differences between conditions. For example, if the fidelity ratings demonstrate poor adherence to the content or process, then non-significant outcomes may be a result of poor implementation rather than an ineffective intervention.

**Inter-rater Reliability** All studies should anticipate calculating a score for inter-rater reliability for each measure to report in publications. The correct measure of inter-rater reliability depends on the type of data, the number of raters and how raters are selected for each session. There are different mathematical formulas to calculate inter-rater agreement and the study design will determine the correct one (Shrout & Fleiss, 1979). More extensive discussion of selecting the type of inter-rater reliability measures can be found elsewhere (e.g., Hallgren, 2012; Landis & Koch, 1977; Shrout & Fleiss, 1979). An important point is that the formula for calculating kappa accounts for agreement by chance and is not just the percent of agreement between two raters. Therefore, a kappa score of 0.66 is different than the raters agreeing 66% of the time. Kappa scores above 0.61 are considered "good" (Altman, 1991) or "substantial" (Landis & Koch, 1977) agreement. For categorical or continuous rating scales, an inter-class correlation (ICC) is the appropriate calculation (Hallgren, 2012). Hallgren (2012) may be a useful resource for determining the correct type of inter-rater reliability for the scale and design of a particular study. However, despite the extensive discussion in the literature of types of inter-rater agreement and the reporting of kappa in intervention studies, the kappa value only establishes the level of agreement reached between two or more raters. It does not convey any information about the degree of fidelity to the model that was achieved in the intervention.

**Fidelity Measurement as Moderator** In the analyses, the percent of content delivered or the average process score can be used to compare across groups or individuals. For interventions with session-specific content, where only a selection of sessions are rated, the fidelity scores for some sub-groups can be compared, such as by site or practitioner. The sample of sessions for each practitioner or site can be used to generalize to all of the clients served by that practitioner or site. If all sessions are rated, the percent content and the process score can be used as an individual-level characteristic and associated with outcomes at the individual level. The only way to include fidelity rat-

ing at the individual level for interventions with session-specific content is to rate all sessions. With a more flexible intervention, where content is not assigned to a session, a sample of each client's sessions could be rated to provide a fidelity score for that client.

**Dose as Moderator** Dose is an individual-level variable that should be recorded for each client. Dose can be used in analyses to determine either a threshold effect (i.e., if participants receive at least half the intervention they show improvements) or a linear dose-response relationship (i.e., there is some improvement for every additional session completed). In a group intervention, where a participant may only attend some sessions, identifying what content each participant received can help identify which components are related to outcomes. For example, if a participant misses a key group session on anger management, he may be less likely to show improvement on anger management than another participant who attended fewer total sessions but did attend the anger management session.

## Pathways Triple P Case Study

### *Inter-rater Reliability*

The same two raters were used consistently across sessions. This is a fully crossed design (Shrout & Fleiss, 1979) and Cohen's kappa was used to calculate kappa. A kappa score was calculated for each session because each session had a different content measure. The kappa values ranged from 0.44 to 0.97 with all but one session scoring over 0.65—an indication that inter rater reliability was in the good or substantial range (Altman, 1991; Landis & Koch, 1977). For the process scores a two-way mixed-effects ICC model assessing for consistency rather than absolute agreement was used. Because the pairs of raters were constant throughout the sessions and were not randomly assigned, an ICC was calculated for each question in each session. Most of the values were in the good-excellent range or reflected systematic disagreement between raters, where one rater was consistently higher or lower than the other (Hallgren, 2012).

### *Fidelity Score in Analyses*

This project has data on content fidelity, process fidelity and dose, all of which can be included as moderators in the analyses. The average content delivered across sessions was 77.2% and the average process score was 7.56 out of a possible score of 10. This means that, on average, about 77% of the content was delivered and that the practitioners followed the PTP process model about 75% of the time.

### *Dose in Analyses*

Dose was used as a continuous independent variable to understand the relationship between the number of sessions completed and the parent and child outcomes.

### **Implications for Research**

Understanding and replicating the circumstances in which an intervention resulted in successful outcomes for clients or patients is critical to replicating the positive outcomes and delivering effective services to clients (Miller & Rollnick, 2014). More transparent and consistent systems of assessing and monitoring fidelity, that are developed using a standardized format, may help agencies reproduce the positive outcomes from efficacy trials. To accomplish this, reports and articles on interventions should provide sufficient details to allow other scientists to replicate the successful conditions. A specific and nuanced understanding of these conditions includes individual- and organizational-level fidelity variables that were monitored, including some assessment of the quality of the delivery of the intervention. Additionally, the measure used to assess fidelity should be available for other researchers and practitioners. Ideally, if the researchers designed the measure, they should briefly describe the system for developing the fidelity measure in the published results. This is similar to a bench scientist reporting the specific cultures and chemicals they used in their experiments.

#### *Individual-level Variables to Report*

To ensure the replicability of an intervention, the literature is consistent in suggesting that the dose, the percent of the material delivered overall and per-session, the extent of client engagement, and the level of fidelity to the process specified in the manual and training should be reported in detail (Proctor et al., 2011). In addition, although these categories require less on-going monitoring, the mode of delivery, and the qualifications and training of the practitioner delivering the intervention, are important to present as well.

A critical, and often overlooked step, is the extent of compliance and the level of detail that was captured in the compliance measure. Simply reporting the kappa or ICC statistic is insufficient. The kappa and ICC are measures of agreement between raters (Shrout & Fleiss, 1979). These statistics do not convey the amount of the intervention delivered. Additionally, the quality and level of detail in the fidelity measure are important. A fidelity measure with only a few general categories that only encompass the main topics of the session provides very little information about how much of the content was delivered. A measure with only a few items may not capture the variation that was present in the amount of content and quality of delivery. To promote

the successful scale-up of interventions, the process of measuring and monitoring fidelity needs to be more transparent. The research community has an obligation to provide more details on how to achieve the positive results presented in journal articles.

#### *Organizational Variables to Report*

In addition to individual level content and process factors, organizational level factors are instrumental in maintaining fidelity. These factors should be monitored and recorded throughout the process. Using extant fidelity frameworks (Cross & West, 2011; Moncher & Prinz, 1991; Schoenwald et al., 2011) other factors to monitor include: (1) the pre-requisites for education, training, and experience for practitioners; (2) the practitioner's training in the intervention, including any certification; (3) the process for supervision; (4) the supervisor's training and experience; (5) on-going training in the intervention (e.g., maintaining active certification through annual continuing education); (6) caseloads; and (7) any other responsibilities that practitioners had in addition to their caseload that may have impacted the amount of time they could spend on service delivery. These pieces of information are important for other researchers interested in replicating the results or for implementation in usual care.

This Field Guide provides guidance on how to develop and implement a comprehensive fidelity rating system. This approach could be applied to other manualized interventions that need a more detailed fidelity rating system than the developers provided so that the fidelity can be rigorously assessed in a research study.

### **Implications for Practice**

For administrators and practitioners, information about the assessment of fidelity should be scrutinized as carefully as the results of the intervention. To achieve the positive results for clients in usual care that were demonstrated in a research study, monitoring and maintaining fidelity must be built into the estimated cost of delivering the intervention. Administrators and practitioners interested in scaling-up the delivery of EBPs should attend to all of the aspects of fidelity. To be competitive for external funding, agencies and organizations are increasingly being asked to demonstrate whether and how they will maintain fidelity to the intervention. This can be accomplished in usual care, but it requires attention, planning, and resources (citation removed for blind review).

While this Field Guide was designed initially to assess fidelity in a research study, it could be used to develop a fidelity rating system for on-going monitoring in the supervision process at an agency.

## Limitations

The *Field Guide* system has a few limitations, as do the measures that were developed for the study. The PTP materials include detailed information about the intended process and content of the intervention in the training and manuals. This provided a good starting point and direction from which to develop the intervention measurement tool. However, the manuals for some interventions may be less specific and it could be more difficult to develop accurate measures for those interventions without additional guidance or details. While the *Field Guide* was developed using the best evidence that was available at the time, and it did produce a useful system for the PTP study, it has not been applied to other interventions. Similarly, while the measures that were developed were useful in our study, they have not been tested in other studies.

## Conclusion

Measuring and maintaining fidelity is critically important in intervention research studies to confirm that the intervention is being delivered as it was designed so that the results can be accurately attributed to the intervention. Fidelity data can then be incorporated into analyses to better understand the role of adherence in treatment outcomes. Given the importance of fidelity in documenting the effectiveness of an intervention, developing a reliable and valid measure of fidelity should receive the same level of attention as other aspects of the study. However, this is often not the case. While the existing literature describes how to build fidelity into a research study (Bellg et al., 2004) and explains the importance of fidelity, the process of developing a system to assess and maintain fidelity, including how to create the specific measures, is notably absent. The five-step process described here for developing a fidelity system for an intervention evaluation study aims to fill this gap and improve the ability of research teams to accurately measure fidelity from multiple perspectives. Measuring fidelity is a critical step in any intervention study and this article fills a gap in the literature by providing detailed step-by-step instructions on developing the system and measures. By systematizing how to develop a fidelity system and measures, we hope to help other researchers avoid common pitfalls and to have complete and useful fidelity data for their analyses. The PTP case study clearly illustrates the process of developing a fidelity measurement system, and identifies the many decisions that need to be made throughout the process.

**Acknowledgements** This research was funded by a grant from the Eunice Kennedy Shriver National Institute of Child Health and

Human Development (1R01HD061454) to Washington University in St. Louis. This work was also supported by the National Institute on Drug Abuse (F31DA034442, K. Seay, PI; 5T32DA015035), a Doris Duke Fellowship, and a Fahs-Beck Doctoral Dissertation Grant. The work was supported by The Washington University Institute of Clinical and Translational Sciences grant UL1 TR000448 from the National Center for Advancing Translational Sciences. TL1 Trainee, subaward TL1 TR000449. Points of view, opinions and content are solely the responsibility of the authors and do not necessarily represent the official views of the NIH, the NICHD or NIDA.

## Compliance with Ethical Standards

**Ethical Approval** All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional research committee, approved by the institutional review board and are in compliance with the 1964 Helsinki declaration and its later amendments.

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

## References

- Aarons, G. A., Sommerfeld, D. H., Hecht, D. B., Silovsky, J. F., & Chaffin, M. J. (2009). The impact of evidence-based practice implementation and fidelity monitoring on staff turnover: Evidence for a protective effect. *Journal of Consulting and Clinical Psychology, 77*(2), 270.
- Altman, D. G. (1991). Some common problems in medical research. *Practical Statistics for Medical Research, 1*, 396–403.
- Beidas, R. S., Benjamin, C. L., Puleo, C. M., Edmunds, J. M., & Kendall, P. C. (2010). Flexible applications of the coping cat program for anxious youth. *Cognitive and Behavioral Practice, 17*(2), 142–153.
- Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., ... Czajkowski, S. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH behavior change consortium. *Health Psychology, 23*(5), 443.
- Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science, 2*(40), 1–9.
- Cross, W., & West, J. (2011). Examining implementer fidelity: Conceptualising and measuring adherence and competence. *Journal of Children's Services, 6*(1), 18–33.
- Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review, 18*(1), 23–45.
- Dobson, D., & Cook, T. J. (1980). Avoiding type III error in program evaluation: Results from a field experiment. *Evaluation and Program Planning, 3*(4), 269–276.
- Dumas, J. E., Lynch, A. M., Laughlin, J. E., Phillips Smith, E., & Prinz, R. J. (2001). Promoting intervention fidelity: Conceptual issues, methods, and preliminary results from the EARLY ALLIANCE prevention trial. *American Journal of Preventive Medicine, 20*(1), 38–47.
- 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.

- Forgatch, M. S., Patterson, G. R., & DeGarmo, D. S. (2006). Evaluating fidelity: Predictive validity for a measure of competent adherence to the Oregon model of parent management training. *Behavior Therapy, 36*(1), 3–13.
- Gearing, R. E., El-Bassel, N., Ghesquiere, A., Baldwin, S., Gillies, J., & Ngeow, E. (2011). Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. *Clinical Psychology Review, 31*(1), 79–88.
- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23.
- Hamilton, J. D., Kendall, P. C., Gosch, E., Furr, J. M., & Sood, E. (2008). Flexibility within fidelity. *Journal of the American Academy of Child & Adolescent Psychiatry, 47*(9), 987–993.
- Hodge, D. R., & Gillespie, D. F. (2007). Phrase completion scales: A better measurement approach than likert scales? *Journal of Social Service Research, 33*(4), 1–12.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.
- Madson, M. B., & Campbell, T. C. (2006). Measures of fidelity in motivational enhancement: A systematic review. *Journal of Substance Abuse Treatment, 31*(1), 67–73.
- Mazzucchelli, T. G., & Sanders, M. R. (2010). Facilitating practitioner flexibility within an empirically supported intervention: Lessons from a system of parenting support. *Clinical Psychology: Science and Practice, 17*(3), 238–252.
- Miller, W. R., & Rollnick, S. (2014). The effectiveness and ineffectiveness of complex behavioral interventions: Impact of treatment fidelity. *Contemporary Clinical Trials, 37*(2), 234–241.
- Moncher, F. J., & Prinz, R. J. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review, 11*(3), 247–266.
- Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., ... Hensley, M. (2011). Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health and Mental Health Services Research, 38*(2), 65–76.
- Sanders, M. R., Cann, W., & Markie-Dadds, C. (2003a). The triple P-positive parenting programme: a universal population-level approach to the prevention of child abuse. *Child Abuse Review, 12*(3), 155–171.
- Sanders, M.R., Markie-Dadds, C., & Turner, K. (2001). *Practitioner's manual for standard triple P*. Brisbane: Triple P International Pty. Ltd.
- Sanders, M.R., Markie-Dadds, C., & Turner, K. (2003b). *Every parent's family workbook*. Brisbane: Triple P International Pty. Ltd.
- Sanders, M. R., & Pidgeon, A. M. (2005). *Practitioner's manual for pathways triple P*. Brisbane: Triple P International Pty. Ltd.
- Sanders, M. R., Pidgeon, A. M., Gravestock, F., Connors, M. D., Brown, S., & Young, R. W. (2004). Does parental attributional retraining and anger management enhance the effects of the triple P-positive parenting program with parents at risk of child maltreatment? *Behavior Therapy, 35*(3), 513–535.
- Sanders, M. R., Turner, K. M. T., & Markie-Dadds, C. (2002). The development and dissemination of the triple P-positive parenting program: A multilevel, evidence-based system of parenting and family support. *Prevention Science, 3*(3), 173–189. doi:[10.1023/A:1019942516231](https://doi.org/10.1023/A:1019942516231).
- Schoenwald, S. K. (2011). It's a bird, it's a plane, it's... fidelity measurement in the real world. *Clinical Psychology: Science and Practice, 18*(2), 142–147.
- Schoenwald, S. K., & Garland, A. F. (2013). A review of treatment adherence measurement methods. *Psychological assessment, 25*(1), 146.
- 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.
- Seay, K. D., Byers, K., Feely, M., Lanier, P. Maguire-Jack, K., & McGill, T. (2015). Scaling up: Replicating promising interventions with fidelity. In D. Daro, A. Cohn Donnelly, L. Huang, & B. Powell (Eds.), *Advances in child abuse prevention knowledge: The perspective of new leadership* (pp. 179–201). New York: Springer.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420.
- Turner, K. M. T., Markie-Dadds, C., & Sanders, M. R. (2002). *Facilitator's manual for group triple P*. Brisbane: Families International Publishing.