Adm Policy Ment Health (2011) 38:32–43
DOI 10.1007/s10488-010-0321-0

ORIGINAL PAPER

Toward the Effective and Efficient Measurement of Implementation Fidelity

Sonja K. Schoenwald · Ann F. Garland · Jason E. Chapman · Stacy L. Frazier · Ashli J. Sheidow · Michael A. Southam-Gerow

Published online: 20 October 2010

Abstract Implementation science in mental health is informed by other academic disciplines and industries. Conceptual and methodological territory charted in psychotherapy research is pertinent to two elements of the conceptual model of implementation posited by Aarons and colleagues (2010)—implementation fidelity and innovation feedback systems. Key characteristics of scientifically validated fidelity instruments, and of the feasibility of their use in routine care, are presented. The challenges of ensuring fidelity measurement methods are both effective (scientifically validated) and efficient (feasible and useful in routine care) are identified as are examples of implementation research attempting to balance these attributes of fidelity measurement.

Keywords Implementation fidelity · Fidelity measurement methods · Adherence

Author Note Sonja K. Schoenwald is a Board Member and stockholder in MST Services, LLC, which has the exclusive licensing agreement through the Medical University of South Carolina for the dissemination of MST technology.

S. K. Schoenwald (✉) · J. E. Chapman · A. J. Sheidow
Family Services Research Center, Department of Psychiatry and Behavioral Sciences, Medical University of South Carolina, 67 President Street, Ste. MC406, MSC 861, Charleston, SC 29425, USA
e-mail: schoensk@musc.edu

A. F. Garland
Child and Adolescent Services Research Center, University of California, San Diego, CA, USA

S. L. Frazier
University of Illinois, Chicago, IL, USA

M. A. Southam-Gerow
Virginia Commonwealth University, Richmond, VA, USA

The conceptual model of implementation described by Aarons and colleagues (Aarons et al. 2010, this issue) focuses on the implementation of evidence-based interventions for children and their families served. Among the implementation process elements identified in the model (see Fig. 1, Aarons et al.) are: "Establish/maintain a clear fidelity focus," and, "Establish innovation monitoring feedback system." These elements reflect the migration into implementation research of constructs originally defined, and to varying degrees, measured, in the psychotherapy treatment outcome literature. The objectives of this article are to highlight key issues in the conceptualization and measurement of fidelity in that literature and their implications for implementation research.

Why Care About Fidelity Measurement?

As noted in organizational research on innovation implementation, innovations can be products or processes; conceptualized more broadly or more narrowly; and designed to be a mostly stable thing to be put into effect, or designed to change while being put into effect. However, regardless of the nature of the innovation "sophisticated, complete or faithful use are always defined normatively according to the inventor's, developer's, or researcher's notion of how the innovation ought to be used to get the best effect (Real and Poole, 2005, p. 76)." The specification of faithful use is the heart of fidelity measurement.

In psychotherapy research, the innovation in question has typically been a particular psychosocial treatment for a

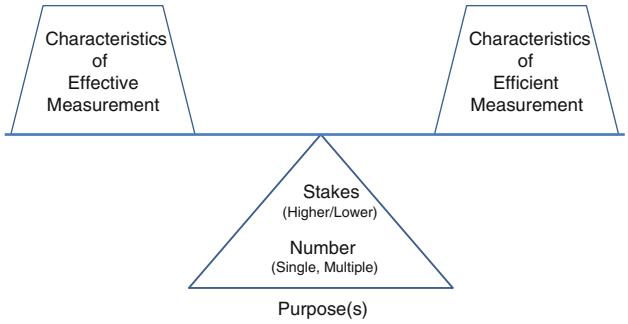

Fig. 1 Balancing efforts to develop effective and efficient fidelity instruments on the basis of purpose

particular disorder or class of disorders (e.g., depression, anxiety disorders) or clinical problem area (e.g., substance abuse, aggression in children). On the basis of research and theory about the correlates of the problem, a treatment theory is developed that identifies mechanisms by which the specific treatment is expected to effect change, and that theory informs the specific contours, content, and processes of the treatment. Since the early 1990s, the development of treatment manuals has made more transparent and therefore replicable a variety of empirically supported, theory-based treatments. Using a manual, however, does not guarantee effective implementation of an intervention (Forgatch et al. 2005; Schoenwald and Henggeler 2004). Instead, intervention delivery must be evaluated for fidelity to content and process so that one can explain whether failure to replicate desired outcomes is a problem with the intervention or of its application (Fixsen et al. 2005; Forgatch et al. 2005; Mihalic 2004). This distinction between intervention and application failure is not unique to mental health, or even to human services in general (Rossi 1978); but has long been documented in research on organizational innovation and the adoption and implementation of innovations in industry and government (Real and Poole 2005).

As demand increases for broader penetration of evidence-based treatments in the services marketplace, so does the expectation that service providers and organizations be held accountable for their outcomes. Whether implementing a particular treatment model or treatment elements common across a class of empirically tested treatments (e.g., cognitive-behavioral treatments, interpersonal treatments; see, e.g., Chorpita and Daleiden 2009; Garland et al. 2008a; McHugh et al. 2009), indices of implementation fidelity are needed to determine whether client improvement or lack thereof is a function of the failure of the treatment (or treatment element) or of its application.

Whereas the origins of fidelity measurement in psychotherapy lay primarily in the need to differentiate an experimental treatment condition from the control condition in efficacy trials, the ubiquity of the "intervention versus application" problem, and the relatively recent recognition that implementation has typically been overlooked in the journey from treatment efficacy to dissemination, conspire to place fidelity measurement squarely in the arena of implementation science. Unfortunately, methods for assessing intervention fidelity are not universally well established, and little is known about the extent to which established methods are being or could be used in routine care settings. Key challenges to achieving adequate measurement of intervention fidelity are reviewed in subsequent sections of this paper, including specific examples of strategies to address these challenges.

A Brief Primer on Fidelity and Its Measurement

There are three components of treatment fidelity: therapist adherence, therapist competence, and treatment differentiation. Therapist treatment *adherence* is the degree to which a therapist uses prescribed procedures and avoids proscribed procedures. Prescribed procedures are those for which use is required to execute the treatment as intended. Proscribed procedures are those for which use is prohibited, either expressly (as would be ideal, but appears rarely to occur in manuals and fidelity instruments), or by virtue of their omission from prescribed procedures, or on the basis of inference from prescribed procedures. For example, manuals for the cognitive behavioral treatment of anxiety may not expressly forbid the practice of dream interpretation, but because dream interpretation techniques are absent from the manual, it is reasonable to infer that dream interpretation is a proscribed element of the treatment. Treatment *differentiation* is the extent to which treatments differ from one another on critical dimensions (Waltz et al. 1993). Therapist *competence* is the level of skill and judgment used in executing the treatment. Each component of integrity captures a unique aspect of treatment integrity that together, and/or in isolation, may be responsible for therapeutic change or lack thereof (Perepletchikova et al. 2007).

Theoretically, adherence, competence, and differentiation could be measured for each component of a treatment, within each treatment session, and overall, as recommended by some experts (Perepletchikova et al. 2007). The specification of the treatment itself should, however, inform the most salient targets of fidelity assessment. For example, a cognitive-behavioral treatment for anxiety may be specified in terms of distinct components (e.g., relaxation training, thought-stopping), while an interpersonal treatment for anxiety may not. Similarly, some treatments are specified on a session-by-session basis, others in terms

of treatment phases with functionally distinct purposes, and others in terms of principles, guidelines, components, or activities that might be used throughout treatment.

Treatment integrity can be assessed using direct or indirect methods. Direct methods are observational, and require a trained observer to view and rate treatment sessions. The sessions are typically video- or audio-taped, although observation of live sessions is sometimes accomplished via the use of one-way mirrors. Indirect methods of integrity assessment include questionnaires or checklists completed by therapists, clients, or experts; review of homework completed by clients; or third party review of written case notes.

Current Status and Knowledge Gaps

In psychotherapy research, treatment adherence has been the more frequently measured aspect of treatment fidelity. Even so, the frequency of its measurement has been relatively low. For example, qualitative reviews of child treatment and family therapy outcome studies conducted over a decade ago revealed that fewer than half of child treatment studies and even fewer family therapy studies included adherence measurement (Hogue et al. 1996). Observational instruments typically specified therapy techniques in molecular terms (e.g., specific verbal behaviors within sessions), while indirect methods of measurement (e.g., checklists, questionnaires) specified constructs in more molar terms (e.g., techniques used across an entire treatment session or component; Heaton et al. 1995). Moreover, the two methods of adherence evaluation rarely correlated even when attempting to index the same construct (Heaton et al. 1995; Waltz et al. 1993). With few exceptions, the same adherence measure was rarely used in more than one study (Malik et al. 2003). There has been some progress in this regard as a result of the larger scale transport and implementation in routine care of some evidence-based treatments. However, a recent review of adult and child psychotherapy outcome studies published between 2000–2004 in high impact journals found only 3.5% of the 147 articles met criteria for adequate methods to establish, assess, evaluate and report measurement of treatment integrity. Although adequate measurement of intervention fidelity was found in only a small minority of studies, most studies did include methods to establish intervention integrity (Perepletchikova et al. 2007). This reflects the reality that methods used to *establish* and *maintain* intervention integrity are not necessarily those used to *measure* and *report* it.

Accordingly, and consistent with earlier reviews (Hogue et al. 1996) the Implementation Methods Research Group (IMRG; see Introduction of this Special Section) has found it helpful to characterize as "quality control" the methods used to establish and maintain intervention integrity. A variety of such methods have been used to ensure an intervention is delivered as intended, including: specification of treatment procedures in manuals; pre-implementation training with or without post-training proficiency criteria; completion of practice cases; treatment model-specific clinical supervision or case consultation; and expert review of clinician treatment notes, session recordings, or self-reports. Such quality control methods indicate the intention to enable clinicians to deliver a particular intervention as intended; absent adequate fidelity measurement, however, the actual delivery of the intervention remains unknown.

An ongoing IMRG review of adherence measurement methods suggests the relative rigor of methods used to ensure and measure treatment integrity can be characterized using a quasi-continuum, depicted in Table 1. Classifying along these dimensions the quality assurance and fidelity measurement methods used in studies of empirically supported psychosocial interventions will facilitate the identification of which types of treatments, for which clinical populations, in which types of settings, have relatively more or less well developed methods to *ensure* integrity and to *measure* integrity. The resulting data will illuminate priorities for research needed to support effective and efficient fidelity measurement in practice contexts.

Table 1 Continuum of quality assurance methods reported in IMRG literature review

Rating	Quality assurance method
1	No report of any quality control methods or fidelity measurement.
2	Report of quality control methods only (e.g., therapist training, specified manual, ongoing supervision/consultation), but no measurement.
3	Report of some fidelity measurement/review, but no specified measure or data reported.
4	Measure of fidelity reported but no data on reliability or validity of measure and/or no test of relationship between fidelity and outcomes.
5	Established measure of fidelity with established psychometrics used and reported and assessment of relationship between fidelity and outcomes reported.

Toward Effective Fidelity Measurement

At first glance, the process of measuring treatment fidelity would seem straightforward: (1) identify relevant treatment components; (2) determine who will provide ratings on the components; (3) obtain ratings on the components; and, (4) devise a summary score based on the ratings. Each step, however, entails many choice points and decisions, the results of which yield very different routes for assessing fidelity, as described next. Crafting instruments that *both* reflect psychometrically sound measurement practices *and* can be implemented in clinical practice presents several challenges. This section describes these challenges and offers a few recommendations for overcoming them.

Establishing the Purpose of a Fidelity Measurement Instrument

The most important consideration in developing a method for measuring treatment fidelity is the purpose of the instrument and intended use of resulting scores. By clearly identifying the exact purpose of the instrument and the intended use of scores resulting from the instrument, subsequent steps deciding on measurement methods will be possible. A simple way to ask this question is, "What decisions do I need to make on the basis of the scores?" The answer to this fundamental question identifies the necessary precision of the scores, and this alone will dictate many of the decisions that follow. For example, a test intended to retain a contract to provide an evidence-based practice or to evaluate employee performance has very high stakes and, thus should be designed to have a very high degree of precision and accuracy. In contrast, a test used in conjunction with other data (e.g., direct observation of sessions) to chart the direction for additional therapist training has lower stakes and may appropriately have lower accuracy. And, while the highly precise test developed for a high-stakes purpose may also turn out to be valid for lower stakes purposes, the costs associated with the former (described subsequently) may render inefficient its use for lower stakes situations. It is very difficult to design a measure that yields valid scores for multiple purposes. Yet, as described in a subsequent section, there is interest among practice and research communities in identifying measures that can serve more than one purpose. Accordingly, Fig. 1 depicts the purpose of fidelity measurement as the fulcrum on which rests efforts to balance the effectiveness and efficiency of fidelity measurement methods (see Fig. 1).

In addition, the selection of a measurement theory to inform test development and validation is crucial, as the assumptions and procedures that follow from each of three extant theories—Classical Test Theory, Generalizability Theory, and Item Response Theory—differ sufficiently to yield different kinds of instruments in response to the same set of stimuli. The *Standards for Educational and Psychological Testing* (SEPT; AERA, APA, and NCME 1999) should be a unifying element for any test development or evaluation efforts, and much of the discussion in this section is based on the SEPT.

Defining Treatment Fidelity

Developing an instrument to index fidelity requires grappling with the defining characteristics of the treatment. Treatments vary with respect to: the generality or specificity of the components of the treatment; necessity of components; timing of components; aspects of the component that indicate fidelity; reference period for each fidelity assessment; and, degree of precision in the resulting scores. The following questions can guide consideration of the defining characteristics of a treatment and measurement of fidelity to it. The answers to them form the basis for determining appropriate methods for measuring treatment fidelity. Table 2 summarizes in broad terms and juxtaposes the characteristics of effective (psychometrically sound) and efficient (feasible) fidelity instruments. Critical details of effective measurement are described next; and feasibility and efficiency issues are elaborated in a subsequent section of this article.

What treatment components need to be evaluated to determine implementation fidelity? With the purpose of the measure in mind, the next important step is to define what aspects of treatment need to be assessed. A key challenge in this step is to determine the level of generality or specificity of the components of the treatment. For example, several detailed items may be needed to assess whether a behavior plan was developed, whereas one or two generally phrased items may adequately index a therapist's use of strength-focused language and praise.

How important is the treatment component for determining fidelity? The treatment components that are identified as indicators of implementation fidelity may or may not be considered "essential." The key question is whether the interventionist could be considered to be adherent if a specific component was not completed. For example, a core element of an ecological treatment pertains to the interactions of an individual with others in the family and surrounding environment. A therapist may not engage a youth's school teacher in treatment, but may still be adherent to an ecological approach if other areas of the ecology are engaged. Thus, intervention with individuals other than the client is essential, but specifically engaging the teacher may not be.

Table 2 How characteristics of effective and efficient measurement methods might map onto one another

Characteristics of effective instruments	Characteristics of efficient instruments
Purpose and theory	Purpose and theory
An effective instrument has a clear primary purpose that guides all steps of development.	Different end users may desire different purposes (e.g., training, quality assurance, secure contract to deliver a specific practice).
An effective instrument is developed through application of a specific measurement theory.	Clinical and administrative utility is apparent.
Definition of adherence	*Definition of adherence*
Aspects of the treatment considered essential are identified. The operations indexed (e.g., behaviors, procedures, techniques, principles) are clear and consistent. The indicator of adherence to items is defined (presence/absence, frequency, etc.) and valid.	Adherence definition must be understood and deemed applicable by end users. If multiple purposes will be served, the definition must make sense for these purposes.
Data collection and scoring	*Data collection and scoring*
Criteria are established for the timing and frequency of data collection and for qualifications of raters (observational measures) or respondents (indirect measures). Method for data collection is specified.	The time, training, expertise, equipment, and materials can be made available within the administrative, supervisory, and documentation practices of an organization. Data collection conforms to ethical and professional norms.
Scoring	*Scoring*
Scores map onto the purpose of the measure.	Scores need to be easily interpretable.

When should the treatment component occur? Some identified components for a treatment need to occur during every treatment interaction, while others might occur only once, only at specific sessions, or only in response to specific events. For example, educating a foster parent on adequate monitoring and supervision of a teenager may only need to occur once, while review and revision of a monitoring and supervision plan may need to recur throughout treatment. And, for a treatment that specifies individualization to each client of some treatment activities, a clinician may determine that monitoring and supervision of a youth is already excellent, so this component is intentionally disregarded.

What aspect of the treatment component indicates fidelity? Given the intended use of the resulting scores and treatment components to be assessed, one must determine the actual indicator of fidelity. Is the construct of interest whether or not the component occurred (i.e., presence/absence), how often the component occurred, how well the component was delivered, or something else? The objective is to determine which construct is more important to assess, but this also may differ by treatment component.

What is the reference period for rating fidelity? The intended use of test scores directs the timeframe for which the interventionist–client interactions are being rated. It could be important to rate each interaction, to rate a specific type of interaction but not every time it occurs, or to rate interactions during a particular window of time. This decision in particular has several practical and resource implications. For example, a treatment defined in terms of distinct phases may require some sampling of fidelity assessment at each phase, but not for each session.

Coding Treatment Fidelity

"Measurement" implies numbers, and fidelity measurement aims to represent observed or reported therapeutic operations in a numeric form. Two issues are paramount here: (1) the accuracy of the coding or rating of therapeutic interactions in accordance with the treatment components; and (2) the coding or rating, ordinal or categorical, must be turned into interval scale measures for its intended use, as described in the Data Scoring section. Regarding the former, observational coding systems with trained raters have been the gold standard method used in treatment efficacy trials because of their potential to provide objective and highly specific information regarding clinicians' within session behavior (Hogue et al. 1996; Mowbray et al. 2003). Even in research contexts, however, developing observational methods for rating therapeutic interactions can be challenging. Considerable time and expense is associated with hiring and training raters, developing a rating protocol, obtaining recordings, generating ratings, checking data, and analyzing and reporting data. Informed consent from clients and therapists is typically required to obtain recordings, which must be physically sent or electronically transmitted from the interventionist to the research unit, which must track recordings due, received, coded or not coded. Throughout the coding process, adherence of coders to coding protocols and high inter-rater reliability must be ensured, and resulting data must be managed and analyzed.

These features of traditional observational coding systems present considerable challenges to their routine use in community practice settings. Another approach to fidelity measurement is therapist or supervisor report via paper-

and-pencil ratings on the use of treatment components during interactions with clients. These clinicians would be considered "non-independent, partially trained raters." They are "non independent" because they are involved in treatment implementation. They are "partially trained" because although implementing treatment, they may not be trained in the intended definition and use of items, rating scales, administration methods, and rating reliability of the measure. Partially trained raters offer greater methodological flexibility and lower expense relative to trained raters. Clinician raters may also be able to rate each session or all treatment components, although the added burden of doing so may be excessive, and strategies to decrease burden such as sampling sessions or components can compromise the ability to generate valid scores. In addition, an important caveat is warranted regarding practitioner ratings of their own behavior: The intended uses of scores derived on the basis of practitioner reports may affect the validity of the data. There is high potential for informant bias in an evaluative scenario in which the intended use of scores could be detrimental to the person doing the rating (e.g., therapist job evaluation, supervisor–therapist relationship, service contract loss to organization).

In some clinical settings and research programs, clients or caregivers of clients rate treatment fidelity. As with practitioners, clients and caregivers are non-independent; their reports may, however, be less biased than practitioner reports. In contrast with practitioners and observational raters, clients and caregivers are entirely untrained in the use of the measure and potentially unknowledgeable about the treatment components and their intended use. Thus, they may not be optimal for identifying the occurrence of the components of treatments, and particularly in detecting the gradations of quality. Among implications of using clients or caregivers as raters is that the rating scale likely will operate differently for trained observers, practitioner raters (partially trained), and untrained client or caregiver raters: trained raters, followed by practitioners, would be expected to discriminate more distinct levels of the quality of implementation of treatment components than would clients or caregivers.

Scoring Treatment Fidelity

The final step in fidelity instrument development involves selection and application of a measurement model; that is, a method for turning ratings into scores. Several challenges arise during this step, each having implications for selection of an appropriate measurement model. First, multiple sources of variance typically contribute to a score, and these should be evaluated or accounted for in the measurement model. These sources include the items in the scale, the clients, the therapists, the interaction of the client with the therapist, the individual providing the ratings, and factors such as the agency in which the intervention is delivered. Other characteristics reflected in scores include the administration or rating method (e.g., audiotapes, videotapes, paper-and-pencil reports, interviewer administration by telephone), the performance of the rating scale (e.g., ceiling or floor effects), the reference window (e.g., reference to a single session versus reference to multiple sessions), and dimensionality in the data (e.g., a scale designed to index one construct actually indexes several). Finally, there may be individual differences in the way raters assign ratings (i.e., tendencies to assign more or less severe ratings) or in the way a single rater assigns ratings over time (e.g., changes in the way an individual rates an item as the rater becomes more knowledgeable about the treatment), such that apparent changes in adherence scores could actually reflect changes in rater, rather than therapist, behavior.

There are additional challenges to producing meaningful scores from the obtained ratings. Missing ratings or responses to items are not uncommon, and decisions about how to handle them have potentially significant consequences, depending upon the intended use of the scores. For instance, not producing a score for a coded tape or an administration with missing item responses creates a serious problem if scores are needed for clinical decision-making for a particular client. Fidelity rating protocols and measurement instruments might also include multiple rating scales (e.g., never to always, or poorly done to well done), with the same or different numbers of response options and different response anchors. Thus, the scoring method needs to accommodate such a feature, addressing the variable range of the ordinal responses as well as the variable spacing between response options on different rating scale constructs. Some items pertaining to fidelity measurement might have a zero-inflated response format, either explicitly or implicitly. For example, treatment components that are not expected to occur during each interaction between the therapist and client may receive the lowest rating of "did not occur," but this rating does not indicate whether or not the component should have occurred. To obtain both types of information, an initial item may index whether or not a component occurred, and a follow-up item the extent to which it occurred.

There is no perfect solution for assessing therapist fidelity to a treatment. However, to increase consumer, practitioner, and payer confidence that the interventions requested were delivered, the contours of sound measurement will have to inform the development and testing of practical tools for adherence measurement that can be used in routine care. Examples of solutions that aim to marry sound measurement and practice context feasibility are

described in the final section of this article. The next section describes aspects of routine care most likely to impact feasibility of adherence measurement in those settings.

Toward Efficient Fidelity Measurement

To effectively monitor the fidelity of intervention implementation in routine care, measurement methods must be feasible and ecologically valid (Hayes 1998; Manderscheid 1998). Even a fidelity instrument with strong established psychometric properties will not be useful for dissemination or implementation if it cannot be used relatively easily in a routine practice setting. This section highlights the need to consider the "contextual fit" of intervention fidelity measurement methods within routine practice settings.

As described in the previous section and summarized in Table 2, procedures for fidelity measurement can differ dramatically along many different dimensions. Differences across these dimensions have significant implications for the resources (e.g., time, money, training, equipment) required to implement the measurement methods. If the resource demands of a particular method greatly exceed the resources available in the intended practice context, odds are low the method will be used. Key feasibility and efficiency challenges to be considered in the development of fidelity measurement methods include: the availability of sufficient financial and professional resources to use the method; the extent to which measurement methods alter established administrative and clinical routines in an organization; and, potential clashes with practice norms regarding the observation and evaluation of psychotherapy, an enterprise for which confidentiality is an ethical imperative and the privacy of psychotherapy may be a treasured practice norm.

With respect to financial, professional, and administrative and clinical routines, the observational measurement methods used in efficacy trials would appear to present the greatest challenges to feasibility and efficiency in routine care. Community stakeholders may not be able (or willing) to carry out observational coding (Schoenwald et al. 2000; Weersing et al. 2002). And, the extent to which observational data collection (e.g., audio or videotaping of practice) fits comfortably within the culture and climate of routine practice settings likely varies. Ethical considerations regarding any attempt to observe mental health practice (Garland et al. 2008b) also may interfere with some routine care providers' willingness to utilize this method. On the other hand, Schoenwald and colleagues (Schoenwald et al. 2008) found the majority of community based mental health clinics serving children used live observation of sessions as part of clinical supervision, and a third employed audio or videotaping, thus suggesting observation of clinical practice is not entirely uncommon in routine care. It is not clear, however, how frequently such observation occurs.

Toward Effective and Efficient Fidelity Measurement in Routine Care

Efforts to marry the principles of effective and efficient fidelity measurement described thus far and summarized in Table 2 are reflected in recent and ongoing research being conducted by the authors and their colleagues. Some examples of these efforts are described next.

Modified observational coding systems. Methodological modifications and technological innovations suggest promise for increasing the feasibility of observational measurement in routine care. In the context of a recently completed randomized trial of different approaches to practitioner training and quality assurance in community clinics, Sheidow and colleagues developed and tested the feasibility and reliability of a hybrid observational strategy (Sheidow et al. 2008). The observational system could provide an efficient method both for supervisors to monitor clinician adherence and for clinicians to conduct self-checks on their implementation. That is, the adherence monitoring system could feasibly be taught to practitioners, and they could rate behaviors on their tapes to obtain a somewhat objective assessment of their performance, including identification of problematic areas. One caveat, however, was that this system was designed for a treatment approach (Contingency Management for substance abuse) that has clearly prescribed steps within each session. Thus, the use of such a system may be less appropriate for more complex or individualized treatments. Within the context of a highly specified treatment, though, there was high feasibility and ease of use, as well as sound reliability and preliminary support for validity. These findings suggest the promise of modified observational coding systems as a feasible fidelity measurement strategy in routine care, at least for some types of treatments. Research is needed, however, to determine the types of treatments for which similar modifications can validly be made, and to estimate the costs of use in routine care.

Live observation of intervention sessions is also being used to support and measure the implementation fidelity of SafeCare throughout child welfare systems in Oklahoma (Aarons et al. 2009). An individual serving as a SafeCare consultant or coach observes home visits, coaches the home visitor and models appropriate behaviors (if needed), and after the session completes a checklist of the intervention content delivered by the home visitor trained in SafeCare. This observation-based checklist has not yet been validated. The fact that live observation, coaching

support, and checklist completion is occurring in field settings statewide, however, holds some promise for the use of observational methods to support and measure fidelity in routine care.

Recent technological advances also portend increases in the feasibility of use of observational methods to both support and measure the implementation of effective interventions in routine care. For example, a web-based system helps monitor the implementation of Keeping Foster Parents Trained and Supported (KEEP; Chamberlain et al. 2008), an empirically validated adaptation of Multidimensional Treatment Foster Care (MTFC; Chamberlain 2003) for foster parents and children. KEEP Programs are provided with a pre-programmed laptop computer with a web cam and remote microphone so that recordings of group sessions with foster parents can be streamed over the internet to a secure KEEP server. This server also holds weekly data from a validated measure of foster parent reports on youth behavior, facilitator ratings of foster parent engagement by facilitators, attendance reports, and demographic information on group participants. The website allows the sender to enter any notes, comments, or questions in relation to the video. The clinical consultant, supervisor, and interventionists are able to use this system in a time efficient manner. Once the trainer has reviewed the recording, the facilitator adherence ratings are entered in the system so fidelity ratings are easily linked to specific sessions. Six US and international agencies implementing KEEP are currently using this web-based system (Chamberlain et al. 2010).

Client report methods. Despite promising advances in the development and use of observational tools within routine care, challenges remain regarding the feasibility and contextual fit of observational coding systems in routine care. Accordingly, the development of potentially more efficient therapist-, client-, or caregiver-report integrity measures represents an important goal for dissemination and implementation research (Fixsen et al. 2005; Mihalic 2004; National Institute of Mental Health 2006). The potential of this approach is exemplified by research conducted with Multisystemic Therapy (MST; Henggeler et al. 2009), for which parent reports of therapist adherence have been linked to short- and long-term outcomes in randomized trials and community based implementation studies (see Schoenwald 2008, for a review of these studies). The adherence measure is used by all organizations in the US and internationally licensed to operate MST programs, attesting to the feasibility of the collection and reporting of brief, caregiver-reported adherence measures. Some of the problems with client or caregiver ratings described in the Effective Measurement section do, however, characterize the MST therapist adherence measure. There are, for example, floor effects on the rating scale, suggesting few caregivers are willing to rate therapists very poorly. And, although caregiver ratings of a particular therapist remain fairly stable over time, suggesting there is no "practice effect" (a potentially positive feature of the measure), this also suggests that caregivers detect little change in adherence throughout treatment.

Hepner and colleagues (Hepner et al. 2009) also have designed a self-report instrument for adult patients to report on the extent to which their treatment incorporated elements of evidence-based cognitive-behavioral therapy and/or interpersonal therapy for depression (Psychotherapy Practice Scale (PPS-Patient version). This instrument was developed on the basis of a review of treatment literature and observational adherence coding systems, expert review of items, and pilot interviews with routine care patients to test item comprehension. Items are designed to be practical and self-explanatory, and are rated on a 7-point scale (e.g., never to always) assessing the frequency of use across the course of treatment (e.g., "My therapist asked me to do things that I enjoy doing between sessions" and, "My therapist and I discussed conversations that I had with other people"). Results of preliminary analyses of data from 420 patients in a large managed behavioral health care organization provide support for the internal consistency and factor structure of the instrument, and show it corresponds only modestly with therapists' self-reports of their primary theoretical orientation. The authors note more work is needed to evaluate the validity of this instrument.

Practitioner report methods. An emerging literature on therapist self-report tools to assess psychotherapy practices also provides potential avenues for the development of fidelity instruments that can be used in routine care. Although not designed to measure adherence to a specific treatment protocol, tools to assess practices (evidence-based treatments and other practices) have been developed by several groups of investigators, and several of these tools have been used with samples of clinicians in routine care. For example, The Therapy Process Checklist (TPC) (Weersing et al. 2002) was developed to assess a comprehensive array of psychotherapeutic practice with children and families. The TPC was developed using rigorous psychometric testing and can be used to assess both broad practice patterns and specific therapeutic strategies used with a particular client. Items were derived to represent three of the most common therapeutic models for youth (e.g., Behavioral: use point or token system; Cognitive: challenge irrational beliefs; Psychodynamic: understand effects of early experience). The TPC was tested initially with a sample of psychologists and psychiatrists recruited from national registries who reported on their practice with "typical" patients, and support for its internal consistency, test–retest reliability and factor structure was obtained with

subsequent samples, including a multidisciplinary sample of 87 therapists in community clinics. The TPC was sensitive to within therapist change based on patient characteristics (Weersing et al. 2002). Baumann and colleagues (Baumann et al. 2006) adapted the TPC by adding items relevant to family therapy and used the adapted version to assess "routine" practice among 77 community-based therapists treating children who had been maltreated. The extent to which respondents needed any particular training or definitional clarification to complete this measure is not known, and more research is needed to evaluate its validity.

One clinician report measure of practice has been integrated into a routine care system throughout the state of Hawaii. The "Monthly Treatment and Progress Summary" (MTPS: Child and Adolescent Mental Health Division 2003) is conceptually based on Chorpita and colleagues' Distillation and Matching Model (Chorpita et al. 2005). The measure requires therapists to endorse the practice elements (e.g., activity scheduling) and target problems (e.g., attention problems, delinquency) addressed in sessions across 1 month of treatment. This tool is one component of a larger system reform effort in Hawaii that includes clinician training on practice elements (including provision of resources), as well as use of treatment process and outcome measurement feedback. The MTPS was designed in collaboration with routine care providers and other stakeholders to support the relevance, feasibility, and comprehension of the measure in the routine care context, and its use for over 5 years across a diverse state mental health system, and adaptation for use by providers in Australia (Bearsley-Smith et al. 2008) attests to its feasibility and adaptability. Preliminary analyses support the internal consistency and logical factor structure of the MTPS, but more research is needed to support its validity and association with treatment process data from other sources and with outcomes trajectories (Orimoto et al. 2009; Young et al. 2007).

Bickman and colleagues (Kelley et al. 2010) also have developed a brief clinician self-report form (The Session Report Form; SRF) to assess treatment content and process at the conclusion of every individual session. The SRF has been tested in a large management behavioral health care organization with over 200 clinicians reporting on more than 7,000 sessions, and preliminary psychometric analyses support its internal consistency. As with the aforementioned measures, further research is warranted to examine the validity of the SRF. The TPC, MTPS, and SRF are examples of ongoing efforts to develop psychometrically sound clinician-report instruments to assess practice in routine care settings, and preliminary evidence attests to the feasibility of their use, although there is limited evidence of their validity and reliability. In addition, little research has examined the relationship between self-report measures of practice processes and observational ratings. In a study of motivational interviewing interventions, therapists self-reported use of a variety of treatment approaches, but observers reported limited use of these same strategies (Carroll and Rounsaville 2007). Similarly, Hurlburt and colleagues (2009) found low concordance between observer ratings of psychotherapeutic clinical strategies and therapist self-reports on a small sample of practice in routine care settings for children with disruptive behavior problems. More research is needed to understand the reasons for these measurement discrepancies, which are likely to include those described previously (e.g., variation in rater competencies to distinguish presence, amount, and quality of a prescribed element; changes in ratings over time; possible social desirability effects in practitioner reports arising from concerns about data use, etc.). Given the limitations of self-report measures, significantly more research is needed to validate these tools and the potential utility of similar tools to index fidelity to specific evidence-based treatments.

When Multiple Uses of Fidelity Data Are Desired: Implications for Instrument Development

In routine practice, different stakeholders in mental health may wish to monitor adherence for a variety of purposes. One set of purposes pertains to the accountability of service providers and organizations for client outcomes. The director of an organization contracted to deliver a particular treatment to a specific target population may wish to use adherence data to demonstrate to service contractors (typically one or more public agencies) the treatment was indeed delivered. A clinical supervisor may wish to monitor adherence to one or more treatments or elements of treatment for the purpose of identifying clinician professional development needs; that supervisor and therapist may wish to monitor adherence with each client with the goal of using adherence feedback to guide treatment planning and speed progress for that client; and, the client may wish to know only that the clinician delivered the treatment he or she requested, irrespective of the treatment provided by the same clinician or program to other clients. Accordingly, in routine care, a desirable characteristic of an adherence instrument or set of instruments may be the extent to which its use for different purposes by different stakeholders can be valid.

As noted earlier and depicted in Fig. 1, however, the many decisions entailed in developing a valid and reliable instrument, generally follow from the clear definition of a paramount, if not single, use of the instrument. An ongoing services trial illustrates one attempt to develop fidelity instruments to meet multiple purposes. The study,

conducted by Atkins and colleagues (Cappella et al. 2008), evaluates the viability and effectiveness of a school mental health service model for children with disruptive behavior disorders in high poverty urban communities. The service model includes a mental health services team composed of community mental health agency providers, family advocates, and key opinion leader teachers whose differentiated intervention activities all coalesce around the goal of children's learning. Through a combination of professional development activities and in vivo classroom support for teachers, and parent groups and home visits for families, the mental health team introduces parents and teachers to evidence-based intervention tools designed to influence the empirical predictors of children's learning.

The investigators aimed to create a set of adherence tools for this multi-component service model that simultaneously (1) enabled the investigative team to measure adherence with rigor and examine statistically relations among adherence indicators and outcomes, and (2) provided partner schools and mental health agencies with clinical tools they could use to sustain implementation of the service components once the research ended. Observational measures were deemed too intrusive and expensive to be sustained in clinics and schools. Instead, a fidelity checklist to be completed by the pertinent service participants was developed for each component of the service model: Teachers completed checklists on every professional development meeting; parents reported on every parent group they attended; and, teachers and parents also reported on (a) the frequency and content of classroom and home support provided by the mental health team, respectively, and (b) their own use of recommended intervention tools and strategies. The interventionists—mental health providers and family advocates—reported on the content and perceived quality of clinical supervision designed to support the service model. To determine data collection frequency, the investigators tried to balance evidence regarding the limitations of individual recall over time, practice effects, and the response burden for teachers, parents, and mental health professionals juggling multiple demands in high stress settings. Ultimately, parents reported in the fall and spring on the frequency and content of support provided by their mental health team during the prior month, and on their own use of specific intervention tools. Teachers reported in the fall and spring on services provided by their mental health team, and reported bimonthly (on average) on their implementation of recommended intervention strategies. Finally, mental health providers and family advocates reported monthly on the content and quality of their most recent supervision session.

As to the second purpose of the measure—to support the continued service delivery in schools and clinics after the research ended - agency directors and school principals considered utilizing adherence data to monitor the activities of their staff in decision making related to salary and advancement (a high stakes purpose). Clinical supervisors and school support staff utilized adherence data to identify professional development needs and to structure data-informed dialogue, feedback, and support for their staff related to work performance and skills acquisition (lower stakes purposes).

A critical caveat, however, is that the validity and reliability of these tools remains to be determined, pending the results of data analyses once the trial has ended. These analyses will reveal the extent to which instruments designed both to index fidelity of implementation and to sustain implementation after research has ended can achieve both goals. Quality improvement research suggests the most effective data sources are those that can be used in the feedback loop for improvement and for evaluation. With few exceptions, fidelity data obtained in psychotherapy research trials have not been used in this way. The goals of measuring and improving fidelity need not be mutually exclusive, however, and the data from this trial may illuminate the extent to which a fidelity measurement approach informed primarily by the feasibility of use in routine care (i.e., efficiency) can also be effective (i.e., validly and reliably measure the service being implemented).

Conclusion

Scant research has evaluated potential solutions to the challenges of marrying effective and efficient fidelity measurement. The examples provided here reflect an implementation research agenda emerging from treatment and services research programs focused on somewhat different clinical problem areas and service sectors (child welfare, education, juvenile justice, mental health, substance abuse) that are focused on the common goal of ensuring effective treatments and services are delivered to youth and families. The agenda calls for the development and testing of effective and efficient fidelity measurement methods. Figure 1 depicts the purpose of fidelity measurement as the fulcrum on which rests efforts to balance characteristics of effective and efficient fidelity measurement, and Table 2 summarizes these characteristics. It will be critical for all stakeholders in mental health to understand the implications of multiple decisions made in pursuit of fidelity instruments that balance effectiveness and efficiency, the resulting strengths and limitations of the measures, and the purposes for which they can legitimately be used.

Acknowledgments The primary support for this manuscript was provided by NIMH research grants 1P30MH074678 (J. Landsverk, PI)

and 1P20MH0784458 (M. Atkins, PI). The authors thank the Implementation Methods Research Group (NIMH grant 1P30MH074678) Executive Committee for helpful reviews of manuscript drafts that lead to the inclusion of Table 2 and Figure 1; Marc Atkins, Bruce Chorpita, and David Henry for their participation in Think Tank discussions pertinent to the manuscript; and, V. Robin Weersing.

References

Aarons, G. A., Sommerfeld, D., 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*, 270–280.

Aarons, G. A., Hurlburt, M., & Horwitz, S. M. (2010). Advancing a conceptual model of evidence-based practice implementation in public service sectors. *Administration and Policy in Mental Health and Mental Health Services Research*.

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Baumann, B. L., Kolko, D. J., Collins, K., & Herschell, A. D. (2006). Understanding practitioners' characteristics and perspectives prior to the dissemination of an evidence-based intervention. *Child Abuse and Neglect, 30*, 771–787.

Bearsley-Smith, C., Sellick, K., Chesters, J., Francis, K., & Gippsland Adolescent Depression Research Group. (2008). Treatment content in child and adolescent mental health services: Development of the treatment recording sheet. *Administration and Policy Mental Health, 35*, 423–435.

Cappella, E., Frazier, S. L., Atkins, M. S., Schoenwald, S. K., & Glisson, C. (2008). Enhancing schools' capacity to support children in poverty: An ecological model of school-based mental health services. *Administration and Policy In Mental Health, 35*, 395–409.

Carroll, K. M., & Rounsaville, B. J. (2007). A vision of the next generation of behavioral therapies research in the addictions. *Addiction, 102*, 850–862.

Chamberlain, P. (2003). *Treating chronic juvenile offenders: Advances made through the Oregon multidimensional treatment foster care model*. Washington, DC: American Psychological Association.

Chamberlain, P., Price, J., Leve, L. D., Laurent, H., Landsverk, J. A., Reid, J. B., et al. (2008). Prevention of behavior problems for children in foster care: Outcomes and mediation effects. *Prevention Science, 9*, 17–27.

Chamberlain, P., Sprengelmeyer, P., Saldana, L., & Padget, C. (2010) Web-based observations to monitor treatment fidelity. Manuscript in preparation.

Child and Adolescent Mental Health Division (2003). *Instructions and codebook for provider monthly summaries*. Honolulu, HI: Hawaii Department of Health Child and Adolescent Mental Health Division. www.hawaii.gov/health/mental-health/camhd/resources/prov-agency/library/pdf/paf/paf-001.pdf. Accessed 23 March 2010.

Chorpita, B. F., & Daleiden, E. L. (2009). Mapping evidence-based treatments for children and adolescents: Application of the distillation and matching model to 615 treatments from 322 randomized trials. *Journal of Consulting and Clinical Psychology, 77*, 566–579.

Chorpita, B. F., Daleiden, E. L., & Weisz, J. R. (2005). Identifying and selecting the common elements of evidence based interventions: A distillation and matching model. *Mental Health Services Research, 7*, 5–20.

Fixsen, D. L, Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature*. Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network (FMHI Publication #231).

Forgatch, M. S., Patterson, G. R., & DeGarmo, D. S. (2005). Evaluating fidelity: Predictive validity for a measure of competent adherence to the Oregon model of parent management training. *Behavior Therapy, 36*, 3–13.

Garland, A. F., Hawley, K. M., Brookman-Frazee, L., & Hurlburt, M. S. (2008a). Identifying common elements of evidence-based psychosocial treatments for children's disruptive behavior disorders. *Journal of the American Academy of Child and Adolescent Psychiatry, 47*, 505–514.

Garland, A. F., McCabe, K. M., & Yeh, M. (2008b). Ethical challenges in practice-based mental health services research: Examples from research with children and families. *Clinical Psychology: Science and Practice, 15*, 118–124.

Hayes, S. C. (1998). Market-driven treatment development. *The Behavior Therapist, 21*, 32–33.

Heaton, K. J., Hill, C. E., & Edwards, L. A. (1995). Comparing molecular and molar methods of judging therapist techniques. *Psychotherapy Research, 5*, 141–153.

Henggeler, S. W., Schoenwald, S. K., Borduin, C. M., Rowland, M. D., & Cunningham, P. B. (2009). *Multisystemic therapy for antisocial behavior in children and adolescents* (2nd ed.). New York: The Guilford Press.

Hepner, K. A., Greenwood, G. L., Azocar, F., Miranda, J., & Burnam, M. A. (2009). Usual care psychotherapy for depression in a large managed behavioral health organization. *Administration and Policy in Mental Health and Mental Health Services Research*. doi:10.1007/s10488-009-0247-6.

Hogue, A., Liddle, H. A., & Rowe, C. (1996). Treatment adherence process research in family therapy: A rationale and some practical guidelines. *Psychotherapy, 33*, 332–345.

Hurlburt, M. S., Garland, A. G., Nguyen, K., & Brookman-Frazee, L. (2009). Child and family therapy process: Concordance of therapist and observational perspectives. *Administration and Policy in Mental Health and Mental Health Services Research, 37*, 230–243.

Kelley, S. D., Vides de Andrade, A. R., Sheffer, E., & Bickman, L. (2010). Exploring the black box: Measuring youth treatment process and progress in usual care. *Administration and Policy in Mental Health and Mental Health Services Research.* doi: 10.1007/s10488-010-0298-8.

Malik, M. L., Beutler, L. E., Alimohamed, S., Gallagher-Thompson, D., & Thompson, L. (2003). Are all cognitive therapies alike? A comparison of cognitive and noncognitive therapy process and implications for the application of empirically supported treatments. *Journal of Consulting and Clinical Psychology, 71*, 150–158.

Manderscheid, R. W. (1998). From many to one: Addressing the crisis of quality in managed behavioral health care at the millennium. *The Journal of Behavioral Health Services & Research, 25*, 233–238.

McHugh, R. K., Murray, H. W., & Barlow, D. H. (2009). Balancing fidelity and adaptation in the dissemination of empirically-supported treatments: The promise of transdiagnostic interventions. *Behaviour Research and Therapy, 47*(11), 946–953.

Mihalic, S. (2004). The importance of implementation fidelity. *Emotional and Behavioral Disorders in Youth, 4*(83–86), 99–105.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24*, 315–340.

National Institute of Mental Health. (2006). *The road ahead: Research partnerships to transform services.* A report by the National Advisory Mental Health Council's Workgroup on Services and Clinical Epidemiology Research. Rockville, MD: National Institutes of Health, Department of Health and Human Services.

Orimoto, T. E., Higa-McMillan, C. K., Mueller, C., & Tolman, R. T. (2009, February). *Organization of therapeutic practices in treatment as usual.* Paper presented at the 22nd annual research conference of the Research and Training Center for Children's Mental Health, Tampa, FL.

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*, 829–841.

Real, K., & Poole, M. S. (2005). Innovation implementation: Conceptualization and measurement in organizational research. *Research in Organizational Change and Development, 15*, 63–134.

Rossi, P. H. (1978). Issues in the evaluation of human services delivery. *Evaluation Quarterly, 2*(4), 573–599.

Schoenwald, S. K. (2008). Toward evidence-based transport of evidence-based treatments: MST as an example. *Journal of Child and Adolescent Substance Abuse, 17*(3), 69–91.

Schoenwald, S. K., & Henggeler, S. W. (2004). A public health perspective on the transport of evidence based practices. *Clinical Science and Practice, 11*, 360–363.

Schoenwald, S. K., Henggeler, S. W., Brondino, M. J., & Rowland, M. D. (2000). Multisystemic therapy: Monitoring treatment fidelity. *Family Process, 39*, 83–103.

Schoenwald, S. K., Chapman, J. E., Kelleher, K., Hoagwood, K. E., Landsverk, J., Stevens, J., et al. (2008). A survey of the infrastructure for children's mental health services: Implications for the implementation of empirically supported treatments (ESTs). *Administration and Policy in Mental Health and Mental Health Services Research, 35*, 84–97.

Sheidow, A. J., Donohue, B. C., Hill, H. H., Henggeler, S. W., & Ford, J. D. (2008). Development of an audio-tape review system for supporting adherence to an evidence-based treatment. *Professional Psychology: Research and Practice, 39*, 553–560.

Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. E. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology, 61*, 620–630.

Weersing, V. R., Weisz, J. R., & Donenberg, G. R. (2002). Development of the therapy procedures checklist: A therapist-report measure of technique use in child and adolescent treatment. *Journal of Clinical Child Psychology, 31*, 168–180.

Young, J., Daleiden, E. L., Chorpita, B. F., Schiffman, S., & Mueller, C. W. (2007). Assessing stability between treatment planning documents in a system of care. *Administration and Policy in Mental Health and Mental Health Services Research, 34*, 530–539.