Published in final edited form as:
Clin Psychol (New York). 2011 June ; 18(2): 142–147. doi:10.1111/j.1468-2850.2011.01245.x.

It's a Bird, It's A Plane, It's … Fidelity Measurement In the Real World

Sonja K. Schoenwald
Medical University of South Carolina

Abstract

In psychotherapy research, fidelity instruments were originally developed as manipulation checks in experimental tests of treatment efficacy. The purposes of fidelity measurement are expanding as consumers, administrators, and payers seek to determine the extent to which the interventions purchased are actually received. Emerging purposes for fidelity measurement are described, as are challenges to developing a single instrument that can adequately meet multiple purposes, and that is both effective (psychometrically sound) and efficient (feasibly used in routine care). Examples are provided of efforts to balance these attributes of fidelity measurement, to measure fidelity at multiple levels of the practice context, and to index and evaluate the effects of additional program parameters on client outcomes in routine care.

Keywords

Implementation fidelity; fidelity measurement; adherence

"I know it when I see it" (Justice Stewart, *Jacobellis v. Ohio*, 478 U.S. 184 (1964) is, according to Wikipedia, "A colloquial expression by which the user attempts to categorize an observable fact or event, although the category is subjective or lacks clearly-defined parameters" ("I know it when I see it," 2010). Supreme Court Justice Potter Stewart penned the phrase in lieu of a precise definition of pornography as presented in a film.

"I know it because I invented it." This paraphrase reflects fidelity as typically conceptualized in research on innovation implementation, namely as within the prerogative of the inventor to define and specify. Regardless of the nature of the innovation (product or process; intended for fixed use or flexible use), "sophisticated, complete or faithful use (sic) 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)."

"It can only be known if it is defined and measured in the following very precise ways." Also a paraphrase, this author's attempt to characterize the multi-faceted definitions of treatment integrity -- adherence, differentiation, competence – and criteria used to discern

Correspondence should be addressed to Sonja K. Schoenwald, Family Services Research Center, Department of Psychiatry & Behavioral Sciences, MUSC, 67 President Street, STE MC406, MSC 861, Charleston, SC 29425. schoensk@musc.edu.

The author 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.

For their contributions to the thinking reflected in this commentary, the author thanks her co-authors on the *in press* manuscript cited therein --Ann Garland, Jason Chapman, Stacy Frazier, Ashli Sheidow, and Michael Southam-Gerow -- as well as Bruce Chorpita, Marc Atkins, and David Henry, who, along with the manuscript authors, helped develop a Think Tank presentation on fidelity measurement presented at the 3rd Annual NIH Conference on Dissemination and Implementation (March of 2010).

the adequacy of their measurement. A recent review suggests reliable and valid instruments to detect each facet of integrity exist for very few treatments (Perepletchikova, Treat, & Kazdin, 2007).

In an era when evidence of comparative effectiveness, accountability for outcomes, and the implications of these for consumer choice increasingly coalesce to inform policy and practice, adequate fidelity measurement is needed to understand what is implemented with clients, and to what effect. A lack of reliable and valid instruments that can be used in routine care settings may slow progress toward the public health goal of making more effective treatments and services more widely available, and contribute to pronouncements of wholesale failure when implementation failure is at fault.

In light of these circumstances, Drs. Bond, Becker, and Drake are to be commended for persisting in the journey toward the development and validation of a low-burden indicator to assess in routine care settings the fidelity of one of the few intervention programs with demonstrated effectiveness in the management of serious and persistent psychiatric illness in adults, Individual Placement and Support (IPS). Given what is at stake with respect to the functioning and well being of adults with serious and persistent mental illness, their families, and the communities in which they reside, ensuring an effective program like IPS is widely and adequately implemented arguably advances the health of the individual and of the public. Absent adequate indicators of IPS fidelity, service systems, practitioners, and consumers will be hard pressed to evaluate the extent to which IPS has been implemented, and thus, the extent to which expected – and unexpected -- outcomes are attributable to the program.

Commentary Purpose and Acknowledgements

The aim of this commentary is to consider the evolving purposes of fidelity measurement, and the challenges inherent in constructing fidelity instruments that can adequately achieve more than one purpose and that are both effective and efficient. The commentary is informed by psychotherapy process and outcomes research, mental health services research, and theory and research on the implementation of innovations. It is also informed by the author's collaboration in the development, validation, and transport of fidelity measurement methods at multiple levels of the clinical context for Multisystemic Therapy (MST; Henggeler, Schoenwald, Borduin, Rowland, & Cunningham, 2009); and, with researchers and clinicians evaluating the adaptation, and implementation of other efficacious or effective treatments for children in community clinic (Schoenwald, Kelleher, Weisz, & the Research Network on Youth Mental Health, 2008) and school (Capella, Frazier, Atkins, Schoenwald, & Glisson, 2008) contexts. Across these domains, common issues arise regarding the purposes, nature, and trade-offs of fidelity measurement methods that can be used in routine care settings. A detailed articulation of these issues appears elsewhere (Schoenwald, Garland, Chapman, Frazier, Sheidow, & Southam-Gerow, in press).

The IPS Fidelity Instrument Signals Challenges to Fidelity Measurement in the Real World

Accurately assessing intervention program content presents both conceptual and methodological challenges. As illustrated in the paper by Bond and colleagues, the assessment of which components are critical to a program may vary by stakeholder (consumer, program developer); and some components not identified on an original instrument may turn out to be critical to program definition, thereby necessitating refinement and re-testing of the fidelity instrument. Such refinement and testing introduces methodological challenges, some of which are evidenced in the paper. For example, when

intervention fidelity is evaluated at the program, rather than client or practitioner level, then obtaining a sufficient sample of programs to adequately evaluate psychometric properties, and relations between the instrument and other variables, such as outcomes, is particularly challenging. When raters using an instrument must be trained in the use of the instrument, or in the intervention program being rated, or both, implications for instrument use are both methodological (even for trained raters, ongoing inter-rater reliability checks are needed; when experts in a program are raters, vigilance against bias is needed) and practical (access to raters trained in the program and/or instrument may be limited; costs are associated with training raters and sustaining their reliability). On the other hand, naïve raters (those trained neither in the use of an instrument nor in the program it assesses) may be unable to provide valid and reliable ratings.

Finally, as the authors acknowledge, it is difficult to achieve adequate scientific rigor to demonstrate and improve the psychometric properties of an instrument once service systems begin to utilize it in uncontrolled evaluations with unidentified respondents and service recipients. When data about an instrument emanate as much from gray literature as from peer-reviewed research, the quality and quantity of information needed to adequately evaluate its robustness can be highly variable. And, even within peer-reviewed, well-controlled evaluations, thresholds for some reliability and validity indicators are not universally met. For example, just over half (60%) of studies evaluating relations between IPS fidelity scores and outcomes supported its predictive validity.

Fidelity Measurement Is Not New; But, Reasons for Doing It May Be

In psychotherapy research, fidelity measurement was historically used as an independent variable check in efficacy or effectiveness trials to demonstrate the extent to which an experimental therapy occurred as intended. The “gold standard” of adherence measurement in this context is observational measurement of therapist-client interactions that provide objective and highly specific information regarding clinicians’ within session behavior (Hogue, Liddle, & Rowe, 1996). In the 1990s, transport began of efficacious school-based prevention programs, community-based interventions for delinquent youth, and for adults with serious and persistent disorders (including IPS). Practice context threats to intervention fidelity and outcomes became apparent, highlighting the need to adequately measure and monitor fidelity in routine care settings. Fidelity assessment began to extend to interventions implemented by teams whose members had differentiated roles (Chamberlain, 2003; Gold et al., 2003) and to other aspects of implementation entailed in deploying the clinical intervention, such as staff type, number, and ratio and program structure (Deci, Santos, Hiott, Schoenwald, & Dias, 1995). Some reviews of the implementation of prevention and intervention programs posited multi-dimensional frameworks to measure program fidelity, and recommended assessment of adherence, quality of delivery, program component differentiation, exposure to the intervention, and participant responsiveness or involvement (Dane & Schneider, 1998; Mihalic, 2004).

The evolving conceptualizations and definitions of fidelity raise important questions about contemporary purposes for fidelity measurement. What is the purpose of a particular fidelity measurement instrument? To what extent can an instrument designed for one purpose, for example, as an independent variable check, legitimately be used for another, for example to hire or fire therapists, or to determine whether funding for a service will be continued?

Measurement Purposes

High stakes purposes—A fidelity instrument used to hire or fire staff, or to make decisions about the continuation of funding for a treatment or service, reflects a “high stakes” purpose. Outside of treatment and services research, examples of high stakes

assessments include college entrance exams, medical licensure exams, exams qualifying attorneys to the bar, and so forth. The results of such tests have considerable immediate and longer-term impact on the life of the test taker; and, potentially –in the case of physician or attorney exams, for instance – on the safety of the public. In such cases, the assessment instruments – tests - are designed to reliably distinguish the aptitudes if not also performance of individuals in specific content and practice areas. The precision of item content, scoring, and gradation in item difficulty of such tests is accordingly quite high and validated with large numbers of test-takers. The massive enterprise of standardized test development, validation, and revision attests to its import in the lives of many students and professionals; as does the controversy about the value of the information such tests provide in the context of other information about the candidate.

Independent variable checks—Reasonably high levels of precision, reliability, and validity also characterize at least some observational measures of treatment fidelity in psychotherapy research, whose purpose was to assess whether the independent variable (experimental treatment) was delivered as intended. At stake was the capacity to make valid inferences from experimental studies of newly developed treatments, such that pronouncements of the effectiveness - or lack thereof – of a particular treatment could legitimately be made.

Emerging purposes—The paper by Bond and colleagues illustrates two emerging purposes of fidelity measurement. One is to enable stakeholders in mental health and substance abuse treatment, a group that includes purchasers, providers, and consumers, to discern the extent to which the program purchased was indeed delivered as intended. This is a “high stakes” purpose: individual practitioners stand to lose jobs, provider organizations, the program, and consumers, the service, based on the results of the IPS fidelity instrument.

A second purpose illustrated in the paper is to inform clinical supervision. There is growing interest in the empirical evaluation of the extent to which clinical supervision contributes to the implementation and outcomes of evidence-based treatments (Schoenwald, Sheidow, & Chapman, 2009), and in using fidelity indicators as part of the supervision process. As portended by Bond and colleagues, however, to inform the clinical supervision process, a fidelity instrument originally specified in terms of the presence of broadly defined elements at the program level will likely require revision to capture aspects of implementation that are more practitioner and client specific.

A third purpose of fidelity measurement is emerging from research on psychosocial treatment as executed in routine practice and the extent to which such treatment resembles evidence-based treatments or elements thereof (Garland, Bickman, & Chorpita, 2010). In this context, efforts are underway to develop low burden, low cost methods to assess the treatment techniques that characterize routine practice, including practitioner and client reported methods. There is, however, scant empirical evidence that these reporters can provide valid and reliable assessments of fidelity. For example, evidence is quite mixed regarding the capability of practitioners to accurately report on their own practices (Beidas & Kendall, 2010).

Fidelity Instrument Effectiveness and Efficiency

Both high-stakes standardized tests and well validated observational fidelity instruments can be characterized as *effective* in that they index accurately what they are intended to assess, with good reliability and validity (construct, concurrent, discriminative and predictive) at a level of precision that distinguishes the performance of individuals, and even the performance of a single individual across multiple cases/clients. They are not, however,

terribly *efficient*: Their development, administration, scoring, and use are often very labor intensive and costly. Of course, not all fidelity instruments developed as independent variable checks are observational; not all specify treatment in highly specific terms; and, not all were developed in efficacy studies.

Fidelity instruments vary considerably with respect to respondent; molar versus molecular specification of the phenomena to be observed; inclusion of proscribed (prohibited) behaviors as well as those prescribed; and, whether adherence is rated as present or absent, in terms of amount (not adherent to very adherent), intensity, extensity, or quality. The specification detail of a fidelity instrument often seems to approximate a "form follows function" rule, such that treatments specified in session-by-session, step-by-step manuals often give rise to instruments that are similarly detailed. Similarly, fidelity instruments for treatments manualized in terms of component parts, classes of therapeutic strategies, and domains of clinical focus often contain fewer and less detailed items.

Fidelity Indicators as Part of an Instrument Panel Guiding Implementation

It does not seem likely a single fidelity instrument can adequately serve the purposes of implementation improvement (via clinician training, clinical supervision, program structure realignment), monitoring, and the high stakes determination that an intervention is delivered with enough quality across enough clients to warrant funding or discontinuation thereof. This is so for at least three reasons: (1) Measurement purpose is the primary driver of the many decisions made during the instrument development process (item content, respondent, response options, scoring metrics, sampling frequency, and so forth), and different purposes thus yield different kinds of instruments. (2) Fidelity at multiple levels of implementation (e.g., clinician, team, program) may turn out both to matter and not to look the same. (3) Where other attributes of an intervention– such as dosage, treatment length, or a common factor such as alliance – are shown empirically to affect either fidelity, or outcomes, or both, then indicators of these attributes will likely also be needed to render well-informed high stakes decisions.

Within the last several years, leading researchers have suggested improving the effectiveness of treatments deployed in routine care will require the broader scale use of implementation and outcomes measurement to support clinical decision-making (Bickman, 2008). Computer-assisted clinical decision support systems that track treatment strategies deployed, client progress, and outcomes are among tools being developed and tested in experimental and field studies (see, e.g., Chorpita, Bernstein, & Daleiden, 2008). One goal of this work is to identify both the therapeutic and clinician feedback strategies that maximize the use and impact of effective therapeutic strategies with specific clients during treatment. Valid, reliable, fidelity instruments could be integrated into such systems.

Like other intervention developers facing the journey of treatment transport and dissemination years ago (see, e.g., Schoenwald & Henggeler, 2003), the approach to fidelity measurement taken to transport MST was to validate instruments that index fidelity at different levels of the clinical context (therapist, supervisor, expert consultant) and empirically evaluate relations among these indicators, and youth outcomes. These instruments are used as part of a quality assurance and improvement package that includes information obtained via processes and products that have not yet been empirically evaluated (e.g., a structured program development and start-up process; a semi-annual program implementation review process and accompanying checklist). Sufficient data from dissemination sites are available periodically to evaluate covariation of validated adherence indicators with those not yet validated empirically, and with other aspects of program performance such as treatment duration, premature termination, and youth outcomes.

Evaluation is ongoing of relations among different aspects of program implementation validated indicators of adherence and outcomes.

Conclusion

If there is any truth to the adage that you get what you measure, then fidelity measurement will likely influence in some way what clients get from practitioners and what payers fund. The many mundane aspects of measurement development and testing rarely hold interest among individuals dedicated to the cause of alleviating human suffering; but, absent greater attention to the development and deployment in routine care of instruments that are both effective at achieving a specified purpose and reasonably efficient, it will be difficult to know whether the antidotes to suffering we work so hard to develop and test are indeed being delivered, and to what effect.

Acknowledgments

The primary support for this manuscript was provided by NIMH research grants 1P30MH074678 (J. Landsverk, PI) and 1P20MH0784458 (M. Atkins, PI), and by the Annie E. Casey Foundation.

References

Beidas RS, Kendall PC. Training therapists in evidence-based practice: A critical review of studies from a systems-contextual perspective. Clinical Psychology: Science and Practice. 2010; 17:1–30. [PubMed: 20877441]

Bickman L. A measurement feedback system (MFS) is necessary to improve mental health outcomes. Journal of the American Academy of Child and Adolescent Psychiatry. 2008; 47:1114–1119.10.1097/CHI.0b013e3181825af8 [PubMed: 20566188]

Capella E, Frazier SL, Atkins MS, Schoenwald SK, Glisson C. An ecological model of school-based mental health services: Enhancing schools' capacity to support children in poverty. Administration and Policy in Mental Health and Mental Health Services Research. 2008; 35:395–409. [PubMed: 18581225]

Chamberlain, P. Treating Chronic Violent Offenders: Advances Made Through the Oregon Multidimensional Treatment Foster Care Model. Washington, DC: American Psychological Association; 2003.

Chorpita BF, Bernstein A, Daleiden EL. The Research Network on Youth Mental Health. Driving with roadmaps and dashboards: Using information resources to structure the decision models in Service organizations. Administration and Policy in Mental Health and Mental Health Services Research. 2008; 35:114–123.10.1007/s10488-007-0151-x [PubMed: 17987376]

Dane AV, Schneider BH. Program integrity in primary and early secondary prevention: Are implementation effects out of control? Clinical Psychology Review. 1998; 18:23–45. [PubMed: 9455622]

Deci PA, Santos AB, Hiott DW, Schoenwald S, Dias JK. Dissemination of Assertive Community Treatment programs. Psychiatric Services. 1995; 46:676–687. [PubMed: 7552557]

Garland AF, Bickman L, Chorpita BV. Change what? Identifying quality improvement targets by investigating usual mental health care. Administration and Policy in Mental Health and Mental Health Services Research. 2010; 37:15–26.10.1007/s10488-010-279-y [PubMed: 20177769]

Gold PB, Meisler N, Santos AB, Keleher J, Becker DR, Knoedler WH, Stormer. The program of assertive community treatment: Implementation and dissemination of an evidence-based model of community-based care for persons with severe and persistent mental illness. [Special series.]. Cognitive and Behavioral Practice. 2003; 10:275–277.

Hogue A, Liddle HA, Rowe C. Treatment adherence process research in family therapy: A rationale and some practical guidelines. Psychotherapy. 1996; 33:332–345.

I know it when I see it (n.d.). In *Wikipedia*. Retrieved September 23, 2010 from http://en.wikipedia.org/wiki/I_know_it_when_I_see_it

Jacobellis V. Ohio, 378 U.S. 184 (1964).

Mihalic S. The importance of implementation fidelity. Emotional and Behavioral Disorders in Youth. 2004; 4:83–86. 99–105.

Perepletchikova F, Treat TA, Kazdin AE. Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. Journal of Consulting and Clinical Psychology. 2007; 75:829–841. [PubMed: 18085901]

Schoenwald SK, Garland AF, Chapman JE, Frazier SL, Sheidow AJ, Southam-Gerow MA. Toward effective and efficient measurement of implementation fidelity. Administration and Policy in Mental Health and Mental Health Services Research. in press.

Schoenwald, SK.; Henggeler, SW., editors. Cognitive and Behavioral Practice. Vol. 10. 2003. Current strategies for moving evidence-based interventions into clinical practice. [Special series]; p. 275-323.

Schoenwald SK, Kelleher K, Weisz JR. the Research Network on Youth Mental Health. Building bridges to evidence-based practice: The MacArthur Foundation Child System and Treatment Enhancement Projects (Child STEPs). Administration and Policy in Mental Health and Mental Health Services Research. 2008; 35:66–72.10.1007/s10488-007-0160-9 [PubMed: 18085433]

Schoenwald SK, Sheidow AJ, Chapman JE. Clinical supervision in treatment transport: Effects on adherence and outcomes. Journal of Consulting and Clinical Psychology. 2009; 77:410–421.10.1037/a0013788 [PubMed: 19485583]