Does what we value make a difference in our assessment of implementation fidelity?

Mary K. Styers

mary@magnoliaconsulting.org

Magnolia Consulting, LLC

Abstract:

Evaluators and researchers are continually tasked with making value decisions in the course of study design. In decisions about implementation fidelity, evaluators and researchers place value on specific observations (e.g., self-report, trained observer ratings) and measurement indicators (e.g., dosage, environment, observed use). Each value judgment can strongly influence how a study's implementation fidelity is conceptualized and analyzed. However, across and within fields, evaluators and researchers tend to hold different values in the conceptualization and use of fidelity. As a consequence of differences in value, evaluators and researchers might not always detect relationships between fidelity and outcomes. Drawing on the literature and experiences from two elementary reading and mathematics program efficacy studies, this paper explores differences in methods for measuring fidelity and calculating fidelity variables and offers recommendations.

> 'Does an intervention work?' is one of the fundamental questions that drives evaluation. Although we all may engage with different approaches, in the end, we want to know whether a program did work, is working, or can work. (p. 199, Century, Rudnick, & Freeman, 2010)

> Part of the problem is that people don't know how to read data, how to sift through it or understand it and that's really a challenge for all of us. This is just an insider conversation, but it affects everyone outside of this club: parents, children, taxpayers and employers. And the stakes have never been higher. We must tell the truth and we must tell it clearly. We cannot communicate in an undecipherable code. (Speech to IES Research Conference, Duncan, 2009, June 8)

In an era of educational accountability and transparency, the value of accurate information on program efficacy and effectiveness is great. Implementation fidelity data offer one pathway to understanding why programs succeed or fail (Dusenbury, Brannigan, Falco, & Hansen, 2003), by providing additional information on how a program is used in the classroom. In evaluations of educational programs, I define implementation fidelity as the degree to which teachers implement a program according to implementation guidelines and expectations. Implementation fidelity helps us to understand how teachers used the program in their classrooms.

Because implementation data provide us with a window into how a program is used in the classroom, results from studies of poorly implemented programs should not be given the same weight as results from studies with a high degree of program implementation. Researchers and evaluators should not expect a program to be effective if teachers did not implement a program according to developer specifications.

Furthermore, the degree of teacher implementation may explain study outcomes. For example, some studies have found that higher fidelity is associated with better outcomes (e.g., O'Donnell, 2008), but others have noted this relationship is not always present (e.g., Borelli, Sepinwall, Ernst, Bellg, Czajkowski, Breger, DeFrancesco, Levesque, Sharp, Ogedegbe, Resnick, & Orwig, 2005). In educational evaluations, the link between fidelity and outcomes is of high

importance. Future users want to know how their implementation of the program will ultimately relate to student achievement.

Despite the fact that information on implementation fidelity is a necessity in understanding program outcomes, past research on fidelity and the relation to outcomes is limited (Berkel, Mauricio, Schoenfelder, & Sander, 2011). For example, in the health behavior literature only 27% of studies examined whether sites implemented programs with fidelity (Borelli et al., 2005) and only 24% of studies on behavioral, social and academic interventions had procedures for documenting fidelity. Furthermore, only 8% of behavioral, social and academic intervention studies looked at the relationship between fidelity and outcomes (Dane & Schneider, 1998).

In light of the significance of fidelity data, researchers and evaluators should take steps to evaluate fidelity using a common set of guidelines. However, the current guidelines for proper implementation fidelity vary. For example, the What Works Clearinghouse (2008) guidelines stipulate that researchers and evaluators self-establish guidelines for measuring fidelity in their interventions and suggest measuring fidelity over the course of the study. Little additional guidance is provided. This paper aims to discuss implementation fidelity and value, specifically how researchers and evaluators choices in the assessment of fidelity can affect how it is conceptualized and understood. This paper will examine the relationships between value and fidelity definitions, fidelity measurement, and fidelity analysis through the lens of past research and experiences from an independent evaluation company. Finally, this paper offers some recommendations for fidelity conceptualization, measurement and analysis.

Which concepts are valued in the definition of fidelity?

A search for "implementation fidelity" or "fidelity measurement" brings up a wide variety of articles from various fields, such as psychology, health education, social intervention, evaluation, etc. Overall, researchers tend to agree that implementation fidelity spans five key areas including, (1)

Adherence: delivery of the program as intended; (2) Dosage: quantity of content received by participants; (3) Quality: effectiveness of program delivery; (4) Participant responsiveness: participant engagement in the program; and (5) Program differentiation: user modifications of the program (Caroll, Patterson, Wood, Booth, Rick & Balain, 2007). Some researchers use different terminology with overlapping ideas, such as (1) Structural: Procedural and Educative; and 2) Instructional: Pedagogical and Student Engagement (Century, Rudnick & Freeman, 2010), wherein "structural" refers to adherence and dosage and "instructional" relates to quality and participant responsiveness. Additionally, in health behavior research, some studies use "treatment integrity" in place of "adherence" (e.g., Borelli et al., 2005).

Many researchers believe that the measurement of fidelity is equivalent to measuring adherence (Century, Rudnick, & Freeman, 2010; O'Donnell, 2008) and most studies only consider adherence in fidelity scores, but measurement of adherence alone misses out on the larger picture (Carroll et al., 2007). By only examining adherence, researchers neglect to consider aspects such as teacher quality and participant responsiveness.

In Magnolia Consulting's educational evaluation research on elementary reading and math programs, we have conceptualized implementation as encompassing adherence, dosage, quality, participant responsiveness and program differentiation. In order to define the essential components of each area, we work with program developers to establish a set of implementation guidelines. These guidelines specify the expected adherence and dosage required by the intervention. We assess teacher quality through trained observer ratings of classroom characteristics that are both common across projects and specific to individual programs. We record student engagement and determine types of differentiation in each classroom.

The area in which we have differed from past research is in the quantification of implementation fidelity. We think of implementation fidelity as multi-faceted and consider the

influence of adherence, dosage, quality and participant responsiveness when relating implementation to outcomes.

As mentioned previously, many studies have studied adherence and the relation to outcomes, but these neglected to consider the impact of other relevant areas, such as dosage and teacher quality. Can we expect students to achieve if they have only received half of their math curriculum? Is it informative to know that a teacher fully covered a program but their quality of teaching was poor? Can we expect students to learn if they are not paying attention and are not engaged? These are the types of questions that have challenged researchers and evaluators in the understanding of implementation fidelity. By only valuing adherence and its relation to outcomes, are we missing out on the larger picture?

Recently, Berkel, Mauricio, Schoenfelder, & Sander (2011) and Caroll et al. (2007) have moved beyond conceptualizing fidelity as adherence and developed models to explain the relationship between the five areas of implementation fidelity discussed previously. For example, Berkel et al. (2011) considered the other four domains as moderators to the relationship between adherence and outcomes and Caroll et al. (2007) examined the additional domains as moderators to the relationship between the intervention and adherence. Despite these attempts to develop new models, we are still left without an understanding of how to measure and analyze these potential moderators.

In the assessment of educational interventions, how do researchers and evaluators incorporate these areas into analyses in a manner that is clear and concise to the general public without complicating interpretation? I suggest the necessity of finding a way to meaningfully incorporate all areas of implementation into our understanding of outcomes. However, before we can reach that point, we need to consider the best methods for fidelity measurement.

Which methods are valued in fidelity measurement?

Once a researcher decides on the critical areas of implementation to measure, the next question often becomes, how will I measure it? Past researchers measured fidelity using observations or self-reports (Mowbray, Holter, Teague & Bybee, 2003; O'Donnell, 2008). However, there are costs and benefits to both approaches. For example, by having a small handful of observations researchers ignore the fact that situations and environments change over time. In contrast, too many observations may make analyses difficult because of an overabundance of information (Mowbray et al., 2003). Differing opinions also exist for self-reports. Some studies have suggested teacher self-reports may be positively biased (Resnicow, Davis, Smith, Lazarus-Yaroch, Baranowksi, Baranowksi, Doyle, & Wang, 1998; Schoenwald et al., 2011) whereas others have found that both self-reports and observations are equally effective (Melde, Esbensen, Tusinski, 2006). One resolution to the potential discrepancy in findings is to collect both data types and to support teacher self-report data with observational data (Dusenbury, Brannigan, Falco, & Hansen, 2003).

In terms of areas measured, Brandon, Taum, Toung, Pottenger & Speitel (2008) suggest it might be most effective to assess adherence or dosage through self-report and quality through outside observers. They suggest that quality is a difficult concept to measure, because it requires a subjective judgment of how well someone is using the program. Additionally, users of the program are biased and are not likely to accurately report how well they are using a program. Once a decision is made on the form of implementation data collection, a follow-up question becomes, what specifically are we looking for?

In determining relevant concepts for measurement, previous studies suggest defining critical components of an intervention through consultation with program developers (Century, Rudnick & Freeman, 2010; Mowbray, Holter, Teague, & Bybee, 2003; O'Donnell, 2008). As a result, indicators of the aforementioned areas of adherence, dosage, quality, participant responsiveness and program differentiation should be defined early on through an analysis of critical components. The

determination of study measures through critical components analyses suggests that each study will be unique in the measurement of fidelity because of the specific nature of programs.

However, one study by Century, Rudnick and Freeman (2010) suggests there could be commonalities across educational evaluations. For example, the following components could be measured across multiple educational interventions: (1) Procedural (e.g., time of unit, content coverage); (2) Educative (e.g., coverage of background, standards); (3) Pedagogical (e.g., teacher encouragement of students, use of different materials); and (4) Student engagement (e.g., student engagement with others, socially and intellectually) (Century, Rudnick & Freeman, 2010). As a result, there might be some be some common areas of fidelity measurement across educational evaluations.

To better understand how fidelity is measured and understood across studies, researchers and evaluators would benefit from seeing the specific variables included in observation protocols and measurements of fidelity. Through these specifications, researchers and evaluators could learn and build on experiences from each other. However, many studies do not offer clear explanations of their methods for fidelity measurement (see Brandon, Taum, Young, Pottenger, & Speitel, 2008).

In our studies of educational programs, we use a combination of self-report and observation data. Teachers complete weekly or monthly logs that address adherence, dosage, differentiation and perceptions of student engagement. Additionally, we send trained observers into the classroom twice a year to assess adherence, quality and student engagement using observation protocols. To protect our self-reports from bias, we take steps to emphasize the importance of honest feedback and stipends being tied to the completion logs independent from content.

In our observations and logs, we commonly look for indicators or critical components across multiple areas (see Table 1 for example categories). Indicators are defined and developed through conversations with program developers and past experiences in educational evaluations. During each observation, trained observers rate the apparent presence of anywhere from 15-30

indicators on a 0 to 3 scale, with a 0 indicating that the teacher did not meet the indicator and a 3 indicating that the teacher fully met the indicator.

Table 1. Common implementation fidelity variables in our observations and logs

Area	Examples of categories	Form
Adherence	• Coverage of expected lessons • Coverage of expected lesson components	Logs, Observations
Dosage	• Days spent using the program each week/month • Time spent using the program each day/week	Logs
Quality	• Teacher understanding of student knowledge • Teacher interactions with students • Use of positive reinforcement strategies • Use of individualized instruction, when necessary • Instructional strategies	Observations
Student engagement	• Student on-task behavior • Student interest and engagement	Logs, Observations
Differentiation	• Modifications to the lesson	Logs, Observations

In order to support teachers in their implementation of educational programs, we coordinate with program developers for teachers to participate in beginning of year and follow-up trainings. Additionally, we monitor teacher questions through the weekly surveys and send them to the product trainer and/or program developer for feedback. From the questions and answers, we create an anonymous Q&A document that is sent to all teachers using the program. This document serves as an extra source of support throughout the entire course of the study.

Previous studies have mentioned the costs and benefits of self-report data, with some researchers suggesting the biased nature (Resnicow, et al., 1998; Schoenwald et al., 2011). We tend to agree with Brandon, Taum, Toung, Pottenger & Speitel (2008) that quality is a difficult thing to measure through self-report. In contrast, adherence, dosage, differentiation and student engagement can all be reported with low levels of bias, provided the reporting individual feels confident in the anonymity of their responses. Additionally, by valuing the use of self-report data, we are able to provide a cost effective way to monitor implementation over the entire study and to keep an open

dialogue with our teachers about program progress. Outside of being in the classroom every day, teacher self-reports might be the most cost effective option for monitoring dosage, differentiation, student engagement, and adherence over the course of the entire year.

Which analyses are valued in the quantification of fidelity?

In relating fidelity to outcomes, evaluators and researchers find ways to quantify fidelity measurements into a single score or multiple scores. Similar to previous thoughts about the conceptualization and methods for fidelity, there appears to be little consensus in how to quantify fidelity (Mowbray et al., 2003), with some suggesting that a universal approach may not be possible due to the specificity involved in studying different programs (O'Donell, 2008).

Some researchers and evaluators used a single measure of implementation fidelity (e.g., Brandon, Taum, Young, Pottenger, & Speitel, 2008; Century, Rudnick & Freeman, 2010; Kalafat, Illback & Sanders, 2007), wherein all of the implementation data is aggregated into a single score using relative weights (Century, Rudnick, & Freeman, 2010), relative teacher rankings (Brandon et al., 2008) or summed unstandardized regression coefficients (Kalafat, Illback & Sanders, 2007).

Other studies used multiple scores to assess fidelity, by creating scores for four different critical component areas (Century, Rudnick, & Freeman, 2010) or having item-level fidelity scores (Bond, Drake, McHugo, Rapp & Whitely, 2009). The conventional belief is that the use of one score may miss the larger picture (Century, Rudnick & Freeman, 2010; Mowbray, Holter, Teague & Bybee, 2003) and knowing which specific components are related to study outcomes is ultimately more meaningful than a single score (Century, Rudnick, & Freeman, 2010).

Furthermore, in the quantification of scores, many studies fail to report the validity or reliability of their implementation fidelity measures (Dusenbury et al., 2003), leading some to call upon the report of reliability and validity information (O'Donnell, 2008).

How do researchers and evaluators offer transparency to readers in their reports of educational evaluations? Regardless of method chosen, single or multiple scores, transparency is paramount. If the reader can understand which areas and critical components are included in the quantification of implementation, then it becomes easier to understand how overall implementation in the study related to outcomes. However, the decision to use a single variable or multiple variables might vary depending on the study and/or what the researcher or evaluator values.

For example, in some of our reading studies, a single measure of fidelity has been negatively related to outcomes and in some reading and math studies, it has been positively related. On rare occasions, relationships between overall fidelity and outcomes have reached significance. The lack of variation in fidelity has oftentimes led us to conclude that all teachers implemented the program with high fidelity. In fact, in most of our studies the majority of the sample had overall fidelity rates above 80%. As a consequence, a higher level of observed implementation may sometimes limit the ability to link fidelity with outcomes.

In considering the ways in which we analyze fidelity, we take steps to ensure the reliability and validity of our measures. For every study, we have multiple raters who assess program fidelity through in-person observations. At the beginning of the study, we discuss the observation protocol and what each component of the protocol will look like in the classroom. After viewing classrooms, we debrief and work together to establish high inter-rater reliability. Once researchers receive item-level scores, we check the validity of our scale and subscales using factor analyses. These analyses often reveal that we can explain 36% to 57% of the variance using one to two factors. Additional information on the reliability and validity of self-report (weekly logs/surveys) and observations are available in Table 2.

Table 2. Description of measures

Curriculum Study	Descriptive	Reliability	Validity

Curriculum Study	Descriptive	Reliability	Validity
Reading	Observation: 29 items measured by presence/absence Log: 30 items self-reported weekly by teachers compared to expectations	Established inter-rater reliability during on-site observations	Observation: One factor solution explained 45% of the variance Log: One factor solution explained 38% of the variance Combined: One factor solution explained 36% of the variance
Math	Observation: 22 items measured by apparent strength of an indicator Log: 8 items self-reported weekly by teachers and compared to expectations	Established inter-rater reliability during videotaped observation discussion	Observation: One factor solution explained 47% of the variance Log: Two factor solution explained 57% of the variance (Factors: Adherence, Dosage) Combined: One factor solution explained 37% of the variance

Note. Specific program names are not disclosed to protect client confidentiality.

In analyzing fidelity, we consider the results from factor analyses and item inter-correlations. From that data and the nature of the study, we decide to create single or multiple scores to represent fidelity. Unfortunately, the relationship between the analytical measures of fidelity and outcomes is never a simple one. In some studies, single scores may offer the best option, because of the strength of factor analyses or reader preferences to use a single score, but comparison of self-report and observation data may reveal two different things (see Table 3).

For example, in one of our math studies, self-report data had a non-significant negative relation to achievement gains whereas observation data had a non-significant positive relation to achievement gains. The difference could be a result of any number of factors, such as differences in indicators across methods, teachers implementing the program more fully on days when evaluators visited, etc. In contrast, for one of our reading studies, both measurement types showed non-significant positive relations to gains. The different findings suggest that by only valuing self-report or log data, evaluators and researchers might be missing out on the larger picture. Both are equally

valid measurements of fidelity and should be considered both together and separately in evaluation designs.

Table 3. Different ways to analyze fidelity in curriculum evaluation studies

Curriculum Study	Single Score	Multiple Scores
Reading	Gain = β_0 + r $\beta_0 = \gamma_{00} + \gamma_{01}$(Self-Report only) + γ_{02}(Observation only) + μ_0 • Log Data only, positive relation to gains, p = 0.89 • Observation Data only, positive relation to gains, p = 0.62 Gain = β_0 + r $\beta_0 = \gamma_{00} + \gamma_{01}$ (Observations + Logs) + μ_0 • Combined Observation and Log data, positive relation to gains, p = .57	Gain = β_0 + r $\beta_0 = \gamma_{00} + \gamma_{01}$(Self-Report only) + γ_{02}(Observation 1) + γ_{03}(Observation 2) + γ_{04}(Observation 3) + μ_0 • Self-report (teacher logs), positive relation to gains p = .32 • Observation Score 1 (Class environment), Negative relation to gains, p = .79 • Observation Score 2, Program specific practices, Negative relation to gains, p = .50 • Observation Score 3 (teacher quality), Positive relationship to gains, p = .19
Math	Gain = β_0 + r $\beta_0 = \gamma_{00} + \gamma_{01}$(Self-Report only) + γ_{02}(Observation only) + μ_0 • Log data only, negatively related to gains, p = .52 • Observation data only, positively related to gains, p = .59 Gain = β_0 + r $\beta_0 = \gamma_{00} + \gamma_{01}$(Combined log and self-report) + μ_0 • Combined observation and log data, small positive relation to gains, p = .97	Gain = β_0 + r $\beta_0 = \gamma_{00} + \gamma_{01}$(Self-Report adherence) + γ_{02}(Self-Report dosage) + γ_{03}(Observation 1) + γ_{04}(Observation 2) + γ_{05}(Observation 3) + μ_0 • Self-report adherence, positive relation to gains, p = .57 • Self-report dosage, negative relation to gains, p = .46 • Observation score 1 (Teacher Quality), negative relation to gains, p = .93 • Observation score 2 (Lesson implementation), negatively related to gains, p = .51 • Observation Score 3 (Student engagement), positively related to gains, p = .33

Note. Specific program names are not disclosed to protect client confidentiality.

As mentioned previously, some studies suggested that implementation is quanitified into several different variables, and these variables are related back to outcomes. In order to examine how the use of multiple domains would make a difference, we created multiple scores for implementation and found that results for relevant domains varied by study (see Table 3). For reading, the observation of teacher quality was positive and the closest to approaching significance. In contrast, the observation of teacher quality was a weaker predictor for our math study, wherein the observation of student engagement showed the best relationship to math achievement gains. The differences in relations between variables and outcomes can be explained by any number of factors, such as differences in critical components between studies, differences in the nature of classroom environments, etc.

Taken together, it is imperative that evaluators collect self-report and observation data in their studies, and do a thorough analysis of the underlying factor structure of implementation data for each study. By only valuing one component of fidelity (i.e., adherence or all five indicators), one approach (i.e., self-report or observation) or one method for measuring fidelity (i.e., single score or multiple scores), we may be missing out on the larger picture.

Summary

Most researchers and evaluators tend to agree that implementation encompasses five main areas (i.e., adherence, dosage, quality, participant responsiveness, program differentiation) (Caroll et al., 2007). However, many studies suggest that the measurement of implementation fidelity is equivalent to measuring one domain (i.e., adherence) (Caroll et al., 2007; Century, Rudnick & Freeman, 2010; O'Donnell, 2008). By recognizing that implementation fidelity is multi-faceted, but neglecting to consider multiple areas in the quantification of implementation fidelity we may not fully appreciate the relation between implementation and outcomes. This notion is supported by previous studies that suggested other implementation fidelity variables might serve as moderators in

the relationship between adherence and outcomes (e.g., Berkel et al., 2011) or intervention and adherence (Caroll et al., 2007).

In the measurement of fidelity, observations and self-reports are both valid methods for measurement with costs and benefits for each method. One possible way to reap the benefits of both methods is to collect teacher self-report and observational data, using observational data to confirm data collected through teacher self report (Dusenbury et al., 2003). Once a decision is made on the data collection method, researchers must decide on the specific areas of implementation to examine. In Table 1, I identified common categories of measurement across each of the five main areas of implementation fidelity and noted the data collection method for each type.

Past researchers have used multiple methods to quantify implementation fidelity, with some choosing single scores (e.g., Brandon et al., 2008; Century, Rudnick & Freeman, 2010; Kalafat, Illback & Sanders, 2007) and others choosing multiple scores (e.g., Century, Rudnick & Freeman, 2010; Bond et al., 2009). Many studies fail to report the reliability and validity data for their scores (see Dusenbury et al., 2003) and some have called upon the report of reliability and validity information (O'Donnell, 2008). In our own evaluations, we have examined both single scores and multiple scores and have considered reliability and validity data for each option. In explaining fidelity data to the general public, transparency is paramount. Readers should be able to clearly understand how researchers or evaluators created a fidelity score or scores. However, even with information on how a score is calculated, researchers and evaluators should present data on the reliability and validity of their indicators. Even when the underlying factor structure is established, the decision by the researcher or evaluator to use one score or several is never a clear one and often entails making some sort of value judgment. It involves asking the questions, What will our audiences value and what does the client value? What information is important?

Recommendations on value in fidelity

Based on findings in the literature and from our own experiences in curriculum evaluation, I offer the following recommendations for fidelity measurement:

1. In the understanding of fidelity, evaluators and researchers need to move beyond valuing adherence as the sole predictor and consider the incorporation of other important implementation variables (e.g., student engagement, teacher quality).
2. Evaluators and researchers should take steps to examine the reliability and validity of fidelity data. Before conducting reliability and validity analyses, both groups may have ideas abouot whether to create multiple or single scores, but additional analyses may offer deeper insight into the relationship between implementation and outcomes. Factor analyses should be conducted to determine the underlying factor structure of fidelity variables for each study.
3. Evaluators and researchers need to "value" fidelity and implementation measurements as multi-faceted and essential in the understanding of outcomes. In descriptions of implementation, evaluators and researchers need to be clear on how fidelity is calculated and what aspects are included or "valued" in the study.
4. Finally, it is important to consider that a "one size fits all" approach to measuring implementation fidelity may not be appropriate. There are commonalities in some areas (e.g., indicators for measurement), but differences in others (e.g., single score or multiple scores, observations or self-reports). While evaluators and researchers value coming up with a universal method, it is also equally important to consider each set of implementation variables in the unique context of each study.

References

Berkel, C., Mauricio, A.M., Schoenfelder, E., & Sandler, I.N. (2011). Putting the pieces together: An integrated model of program implementation. *Prevention Science, 12*, 23-33. doi: 10.1007/s11121-010-0186-1

Bond, G.R., Drake, R.E., McHugo, G.J., Rapp, C.A. & Whitley, R. (2009). Strategies for improving fidelity in the national evidence-based practices project. *Research on Social Work Practice, 19*, 569-581. doi: 10.1177/1049731509335531

Borrelli, B., Sepinwall, D., Ernst, D., Bellg, A.J., Czajkowski, S., Breger, R., DeFrancesco, C., Levesque, C., Sharp, D.L., Ogedegbe, G., Resnick, B. & Orwig, D. (2005). A new tool to assess treatment fidelity and evaluation of treatment fidelity across 10 years of health behavior research. *Journal of Consulting and Clinical Psychology, 73*, 852-860. doi: 10.1037/0022-006X.73.5.852

Brandon, P.R., Taum, A.K.H., Young, D.B., Pottenger, F.M., & Speitel, T.W. (2008). The complexity of measuring the quality of program implementation with observations: The case of middle school inquiry based science. *American Journal of Evaluation, 29*, 235-250. doi: 10.1177/1098214008319175.

Burns, M.K., Peters, R. & Noell, G.H. (2008). Using performance feedback to enhance implementation fidelity of the problem-solving team process. *Journal of School Psychology, 46*, 537-550. doi: 10.1016/j.jsp.2008.04.001

Caroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. Implementation Science, 2, 1-9. doi: 10.1186/1748-5908-2-40.

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation, 31*, 199-218. doi: 10.1177/1098214010366173

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*, 23-45.

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: Theory and Practice, *18*, 237-256.

Duncan, A. (2009, June 8). *Robust Data Gives Us The Roadmap to Reform* [Transcript of Address to the Fourth Annual Institute of IES Research Conference]. Retrieved from http://www2.ed.gov/news/speeches/2009/06/06082009.html

Kalafat, J., Illback, R.J., & Sanders, D. (2007). The relationship between implementation fidelity and educational outcomes in a school-based family support program: Development of a model for evaluating multidimensional full-service programs. *Evaluation and Program Planning, 30*, 136-148. doi: 10.1016/j.evalprogplan.2007.01.004

Melde, C., Esbensen, F., Tusinski, K. (2006). Addressing program fidelity using onsite observations and program provider descriptions of program delivery. *Evaluation Review, 30*, 714-740. doi: 10.1177/0193841X06293412

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. doi: 10.1177/109821400302400303

O'Donnell, C.L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research, 78*, 33-84. doi: 10.3102/0034654307313793

Resnicow, K., Davis, M., Smith, M., Lazarus-Yaroch, A., Baranowksi, T., Baranowski, J., Doyle, C.,

& Wang, D.T. (1998). How best to measure implementation of school health curricula: a comparison of three measures. Health Education Research: Theory and Practice, 13, 239-250.

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, 38*, 32-43. doi: 10.1007/s10488-010-0321-0

What Works Clearinghouse (2008). *What Works Clearinghouse: Procedures and Standards Handbook* (Version 2.1). Retrieved from http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=19