TREATMENT FIDELITY IN STUDIES OF EDUCATIONAL INTERVENTION

Edited by Greg Roberts, Sharon Vaughn, S. Natasha Beretvas, and Vivian Wong
## CONTENTS

| List of Figures   | vii         |
| List of Tables   | viii        |
| List of Exhibits | ix          |
| Acknowledgments  | x           |

1 Implementation Fidelity and Educational Science: An Introduction  
   Greg Roberts  

2 Features of Fidelity in Schools and Classrooms: Constructs and Measurement  
   Frank M. Gresham  

3 A Framework for Incorporating Intervention Fidelity in Educational Evaluation Studies  
   William M. Murrah, Jeff Kosovich, and Chris Hulleman  

4 Implementation Fidelity: Examples from the Reading for Understanding Initiative  
   Greg Roberts, Nancy Scammacca Lewis, Anna Mari Fall, and Sharon Vaughn  

5 Intent-to-Treat and Treatment Take-Up Effects of Summer School on the Literacy Outcomes of Struggling Early Readers  
   Keith Zvoch
Contents

6 Linking Implementation Fidelity to Impacts in an RCT 100
   Fatih Unlu, Laurie Bozzi, Carolyn Layzer, Arthur Smith, Cristofer Price, and Richard Hurtig

7 Measuring Treatment Fidelity with Reliability and Validity Across a Program of Intervention Research: Practical and Theoretical Considerations 125
   Jill H. Allor and Lynne Stokes

Technical Appendix 143
List of Contributors 149
Index 150
3

A FRAMEWORK FOR INCORPORATING INTERVENTION FIDELITY IN EDUCATIONAL EVALUATION STUDIES

William M. Murrah, Jeff Kosovich, and Chris Hulleman

The randomized controlled trial (RCT) is considered by many researchers to be the gold standard for rigorously evaluating the efficacy and effectiveness of interventions (IES, 2014). While the RCT is a crucial tool for understanding the average causal impact of an educational intervention, it is less useful for understanding the processes by which an intervention impacts outcomes. In fact, many have described the information about causal effects provided by randomized experiments as a “black box” (Cook, 2003; Imai et al., 2011). In order to open this black box there is growing interest in collecting data on intervention fidelity to inform our understanding of how, and under what conditions, interventions work (Bloom, 2005).

Assessing intervention fidelity will help researchers understand and interpret impact estimates, as well as assess the generalizability of these results to other conditions in which the treatment is implemented. However, if fidelity measures are added to evaluation studies ad hoc and without careful planning, they may fail to capture information about the implementation of the intervention under investigation and the nature of the treatment contrast in study conditions. These measures may not be psychometrically reliable and valid to provide interpretable information. Another concern is that without carefully planned assessment of fidelity, researchers may not be able to evaluate and further develop the conceptual models of the intervention. The five-step framework helps address these concerns by guiding the evaluator in assessing intervention fidelity, starting in the early stages of planning an intervention evaluation and continuing through analyses of results.

Although funding agencies and education researchers have generally come to accept the importance of both assessing intervention fidelity and linking fidelity
assessments to outcomes of interest (IES 2014), there is little information available to educational researchers about the appropriate procedures for accomplishing these goals (Nelson et al., 2012; O’Donnell, 2008; Weiss et al., 2014). The purpose of this chapter is to present a framework to help researchers conceptualize, measure, and analyze the fidelity of specific interventions. First, we describe the framework conceptually, explaining each step. We then give a more detailed example of an intervention evaluation using the framework to give readers a more nuanced understanding of how this framework can be used.

**Five-Step Model of Intervention Fidelity**

The five-step framework focuses on explicating the intervention by breaking it into its primary processes and measuring those processes effectively to evaluate intervention fidelity. As such, it provides a useful guide for educational researchers to consider when designing their RCTs in a way that is sensitive to measurement of fidelity. To be most useful, educational researchers must begin thinking about intervention fidelity early in the development of the intervention study. For more detailed descriptions and discussion of the five-step fidelity assessment framework, we refer readers to the writings of Hulleman and colleagues (Hulleman et al., 2013; Nelson et al., 2012). Below, we offer a brief description of the framework followed by a relevant example.

**Step One: Define the Intervention Logic Models**

Without specifying what is happening inside the black box of the intervention, researchers and evaluators are left wondering what mechanisms explain a successful intervention, or what factors contributed to a failure. As presented in Figure 3.1, unpacking the black box involves specifying two types of processes: intervention and psychological. **Intervention processes** are the core components of the intervention theorized to drive changes in participants and thus impact key psychological processes within the participants. **Psychological processes** refer to the proximal changes in participants, such as depth of knowledge and attitudes that lead to the desired outcome, such as increased learning and achievement. Often, changes in the psychological processes are the goal of the intervention. While identification of outcomes as proximal and distal may depend on the particular analysis being considered, for clarity in this chapter we consider psychological processes as proximal outcomes, and all other outcomes impacted by the proximal outcomes as distal outcomes. Researchers can begin by creating a logic model, which is a graphical depiction—similar to a path diagram—of the core components of the intervention (Nelson et al., 2012). The change logic model is a conceptual representation of the intervention organized in the hypothesized causal order of events. This model should include the intervention core components, psychological processes and intervention outcomes. Once defined, the change model aids researchers in developing an operational logic model that identifies which indicators the researchers
deem important for measuring the core components (Knowlton & Phillips, 2009; Nelson et al., 2012). Development of the operational model involves the specification of each core component and associated subcomponents.

For example, in a randomized experiment of parents by Harackiewicz et al. (2012), parents were targeted with an intervention designed to increase their high school student’s course-taking in science and mathematics. As presented in Figure 3.2, the intervention contained two core components: providing information about the importance of Science, Technology, Engineering, and Mathematics (STEM), and providing guidance on how to talk to teens about STEM topics. There were also two psychological processes targeted: perceived value of STEM to their teens’ lives, and the number of conversations about the value of STEM. The information about the value of STEM was theorized to increase parents' perception of the value of STEM for their teens, and together with guidance on how to talk to their teens, was theorized to increase the conversations that parents had with teens about the value of STEM. This would in turn increase course-taking in high school. Figure 3.3 presents the operational logic model for the parent intervention. The boxes represent the measures or information collected for each of the hypothesized core intervention components, as well as proximal and distal...
FIGURE 3.3 Operational Logic Model for the Parent Intervention (Study 1).

Note: Boxes represent components to be measured. See Kopp et al. (2012) for more information.
outcomes. Thus, the operational model clarifies which aspects are important to measure and what types of measures need to be developed in order to understand the level of intervention fidelity. It is crucial to capture these elements in both the treatment and control conditions so that actualized treatment strength can be ascertained (Hulleman & Cordray, 2009).

Logic models are meant to guide researchers, implementers, and evaluators towards the ideal version of the program of interest. The models specify critical aspects that define each condition (i.e. control, treatment, alternative treatment) as well as the common aspects that span conditions. When theory-based interventions are developed, they are thought to work through very specific mechanisms. Fidelity assessment is meant to measure variation in those mechanisms within and between different groups. A point to note is that aspects of the control group or alternative treatment group may promote psychological processes targeted by one or more core components of the intervention. This could result in attenuation of expected treatment effects related to these components. This possibility highlights the importance of measuring fidelity in the control group as well as the treatment group (Cordray & Pion, 1993).

**Step Two: Identify Fidelity Measures**

After specifying core components and the operational logic model, it is necessary to compile measures of each component (Nelson et al., 2012). The change and operational models provide researchers with a map and an organized inventory of what needs to be measured during the intervention. Direct or indirect measures should be obtained or developed for each component. Each fidelity indicator can be measured through observational data (e.g. live, video-recording, permanent product), self-report data (from implementers or participants), data logs, or any number of other instruments. These measures need not be limited to a single level of setting. In fact, fidelity may operate at every level of an institution; students, teachers, administrators, and schools are all candidates for fidelity measurement. Including this step allows researchers to keep in mind those measures that are necessary for their study, and in which groups they should be measured. For example, if researchers suspect that business-as-usual practices may expose participants to core components of the intervention, they should plan to measure these core components in both groups. As identified in Figure 3.3, fidelity measures in this study included using computer logs (information about importance of STEM), self-report surveys (increased parent perceived utility value of STEM), interviews (increased conversations), and academic transcripts (increased STEM course enrollment in 11th and 12th grade).

**Step Three: Conduct Psychometric Analyses**

Once measures are identified—but before being incorporated within the analysis of treatment variation—dimensionality and reliability evidence should be gathered
for the different fidelity measures (Hulleman et al., 2013; Kosovich, 2013; Nelson et al., 2012). If the decisions related to how to combine fidelity indicators is not consistent with their dimensionality, analyses including these measures can impact the validity of results. Researchers must first determine how the individual fidelity indicators will be combined into scales. Will there be a single, overall measure of fidelity, where all indicators contribute to one index? Or, instead, will there be sub-scales of fidelity that correspond to the core components and subcomponents, to the measures (e.g. self-report, observation), or based on the dimension of fidelity (i.e. exposure, adherence, quality, responsiveness)? The choice of combined indicators should be based on both theoretical and empirical grounds. These combinations should be useful to the goals of the fidelity assessment, but should also reflect the empirical relationships between the indicators. Empirical methods, such as factor analysis, can help evaluators decide how best to combine items (Abry et al., 2012; Brown, 2006; Hulleman et al., 2013), but a strong conceptual framework about how the items should be grouped is essential (Marsh & Hau, 2007). Loss of information and the measurement scale of the indicator should also be considered when combining indicators. Combining indicators typically results in a loss of detailed information; specifically, when combining indicators into a scale, it is common to either average or sum the indicators to create a total composite score. If some indicators are more important than others to the construct, then these distinctions are lost because all indicators are equally weighted. In addition, combining indicators can be challenging because of different metrics. For example, it would be improper to average a binary (0, 1) indicator and a continuous indicator from a nine-point scale because the two indicators would be contributing drastically different amounts of information to the composite.

Once the appropriate dimensions of fidelity have been determined, then the researcher can focus on scale reliability. As with any measure, the use of instruments that produce reliable and valid scores is absolutely critical to generating useful data. The importance of reliability cannot be overstated; good reliability increases the confidence that scores produced are consistent and replicable (Traub & Rowley, 1991). Poor reliability of fidelity measures can lead to bias in models that include fidelity measures as covariates or mediators. Without good reliability, the scores produced cannot be interpreted as accurate representations of focal constructs.

**Step Four: Conduct within- and between-Group Fidelity Analyses**

If the fidelity measures have been deemed psychometrically acceptable, researchers can then begin the fidelity analyses. Within-group descriptive analyses involve investigating the variation in fidelity within each experimental condition. These analyses give researchers a rich description of how well the intervention was implemented as well as how implementation may have varied across individuals and groups (e.g. classrooms) within the experimental conditions. For example,
fidelity can be calculated by computing the average value of a particular fidelity measure, or by computing the proportion of fidelity relative to the total possible fidelity on that measure. Once the basic descriptive statistics have been examined, researchers can then conduct more sophisticated analyses that contrast the treatment and control groups. Such between-group contrasts can be conducted through the use of an Achieved Relative Strength index (ARS) (Hulleman & Cordray, 2009; Hulleman et al., 2013). ARS values are useful in a number of different ways, but are primarily used for comparing a group’s level of measured fidelity to other groups or some absolute standard. For example, we might expect students who receive a value intervention to experience more exposure to value concepts than students who don’t. Alternatively, we might be interested in how far the group is from some ideal standard (e.g. a maximum fidelity value or an important benchmark). The specific index of fidelity should depend on the research or substantive question at hand.

**Achieved Relative Strength Indices**

Several different methods are available for calculating ARS. The easiest method is to compare average differences between components. Another method involves comparing binary complier indices, which can be calculated based on a cut score chosen by the researcher. Finally, researchers can calculate absolute fidelity indices, which determine the achieved proportion of total possible fidelity to a component (i.e. what proportion of participants reached maximum fidelity for a component). Comparing ARS values allow researchers to determine the treatment strength of an intervention. For example, in the parent intervention teenagers reported having more conversations with their parents about STEM than those in the control condition in both 10th grade ($d = .26$) and 11th grade ($d = .39$) (Kopp et al., 2012). In this example, the intervention effect was expected to be stronger in 11th than 10th grade because the second intervention dose was not delivered until early fall of 11th grade. The within- and between-group fidelity analyses are the foundation of intervention fidelity assessment. This step in the framework allows researchers to determine whether or not their intervention was implemented at an acceptable level of fidelity.

**Step Five: Link Fidelity to Outcomes**

As has been demonstrated, a significant amount of work can be done to analyze fidelity without investigating the relationship to outcomes. However, step five is critical for fidelity assessment to be used to its full potential—it allows a researcher to determine if fidelity matters for an intervention. Theoretically, the impact of an intervention on outcomes should increase as fidelity increases and result in appropriate correlations between fidelity measures and outcomes. Linking indices to outcomes can provide several pieces of useful information, including whether
the measured components actually affect outcomes in the expected manner, whether fidelity mediates the effect of the intervention (MacKinnon et al., 2002), or whether there are differences in which intervention components are important for groups characterized by demographic factors (i.e., moderated mediation). Fidelity indices could be added into the statistical model used to predict program outcomes (e.g., Abry et al., 2014; Abry et al., 2012; Kopp et al., 2012). For example, in the parent intervention, a series of logistic regressions could be used to examine the relationship between each of the fidelity indices and both proximal (number of conversations between parents and teens) and distal outcomes (number of STEM courses taken). It may be possible that participants with high levels of incoming value are more likely to exhibit high levels of fidelity. This type of analysis can provide additional diagnostic information for determining how effective an intervention may be in the future. In this way some interventions may be able to identify risk factors for participant noncompliance. However, researchers must be careful in how they interpret analyses that include measures of fidelity. Because fidelity indices are typically collected after random assignment, the strong causal interpretation afforded by randomization requires much more stringent assumptions. Instead, researchers may wish to use fidelity measures in exploratory analyses that complement the intent-to-treat analyses.

**Summary**

The five-step fidelity assessment framework provides a useful and systematic way to assess intervention fidelity in educational settings. It also helps to clarify the conceptual model of an intervention, and allows researchers to organize and identify indicators needed to measure fidelity. The framework also emphasizes the investigation of reliability and validity information to maximize the usefulness of measures. The common need to combine indices further emphasizes the need for good measurement practice and psychometric techniques. Finally, if the first four steps have been effectively executed, the fifth step allows researchers to test whether or not fidelity affects outcomes. As a result, the five-step framework combines theoretical, methodological, and analytical approaches to bolster the validity arguments made about an intervention's effectiveness. We present a summary of the five-step model in Table 3.1.

**Applying the Five-Step Model**

The example in this section uses data from a motivation intervention designed to aid student success in classrooms. This intervention helps students see the relevance and usefulness of the topic they are learning (i.e., utility value) (Eccles et al., 1983) for their current or future goals (Hulleman & Harackiewicz, 2009). One implementation of the intervention delivered in a laboratory (Study 1) (Hulleman et al., 2010) with college undergraduates yielded a significant main effect of
condition—participants in the treatment group reported significantly higher perceptions of utility value than participants in the control group ($\beta = .19, p = .05$). However, another implementation of the intervention delivered in a classroom (Hulleman & Harackiewicz, 2009) produced a much weaker effect—there was almost no difference between the control and treatment groups ($\beta = .08, p = .12$). Whereas student perceptions of utility value increased dramatically in the lab study (Hulleman et al., 2010), increases in perceptions of utility value in the field were minimal (Hulleman & Harackiewicz, 2009). Without measuring fidelity (i.e., intervention processes), it was difficult to conclude whether the intervention worked or not because evidence supported both conclusions. In fact, post hoc investigation of intervention fidelity in the two studies (Hulleman et al., 2010; Hulleman & Harackiewicz, 2009), revealed that fidelity to the intended design was 58% for the lab study and 25% for the field study (Hulleman & Cordray, 2009). These differences corresponded with differences in outcome where the utility intervention had a larger effect in the lab ($g = 0.45$) than in the field.

### TABLE 3.1 The Five-Step Framework of Fidelity Assessment

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
<th>Key Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Define the intervention logic models</td>
<td>Users describe the theory of change through which the intervention is expected to operate. This includes all of the core and common components intervention.</td>
<td>Intervention processes, Psychological processes, Change logic model, Operational logic model, Core intervention components, Common components of educational settings</td>
</tr>
<tr>
<td>2. Identify fidelity measures</td>
<td>Users identify one or more measures for each of the core and common components.</td>
<td></td>
</tr>
<tr>
<td>3. Conduct psychometric analyses</td>
<td>Users demonstrate that data collected is consistent and meaningful for final analyses.</td>
<td>Dimensionality, Reliability</td>
</tr>
<tr>
<td>4. Conduct within- and between-group fidelity analyses</td>
<td>Users determine if fidelity levels for each group are adequate and compare levels across groups to determine if they are sufficiently different from one another.</td>
<td>Achieved relative strength (ARS)</td>
</tr>
<tr>
<td>5. Link fidelity to outcomes</td>
<td>Users determine if variation in fidelity is related to variation in outcomes—typically higher fidelity should lead to more desirable outcomes.</td>
<td></td>
</tr>
</tbody>
</table>
These findings suggest that degradation in fidelity is one possible reason results in the field interventions were weaker. In the following example, we use the five-step framework to evaluate an online version of the utility intervention just described. This replication was implemented as a randomized field experiment in an introductory psychology course over a 15-week semester (complete details about this study, including analyses of outcomes, can be found in Hulleman et al., under review). Students were randomly assigned to one of two different conditions: control (C) or treatment (T). Each student was asked to write two essays (one during week 4 and one during week 8) based on their experimental condition. As reported in Hulleman et al. (under review), however, the treatment yielded small effect sizes.

How Well Was the Intervention Implemented?

The fact that this intervention was conducted within an experimental framework means that fidelity measures can be compared within and across conditions by assessing treatment strength (Hulleman & Cordray, 2009). Fidelity indices are used to calculate achieved relative strength (ARS), which provides researchers with useful information about how effectively an intervention is implemented.

Sample and Procedure

As reported in Hulleman et al. (under review), the intervention was conducted in an introductory psychology course during a 15-week semester. The sample included 206 students. Eligible students completed both surveys and were 18 years old or older (N = 298; 74.2% Female, 61% freshman, 84% non-psychology majors, 84% White, 5% African American). The mean age of participants was 18.7 years. Data were collected throughout the semester. A self-reported measure of utility value (the key psychological process of the intervention) was administered before, during, and after completion of the intervention. During the second week of the class, participants filled out a motivation survey (Time 1 motivation). During the fourth week participants took the first exam (Exam 1). The first intervention essay was assigned after students completed Exam 1, and the second essay was assigned during the eighth week of the semester. During the fourteenth week of the semester, students completed another motivation survey (Time 2 motivation). Finally, during the fifteenth week, students completed their final exams.

Intervention Essays

Fidelity data were collected from student essays and self-report responses and were aggregated across the two intervention essays. In both essays, students were asked to respond to writing prompts that focused on the relevance of course material to their lives (utility value condition) or a summary of course material that they were
currently learning (control condition). Specifically, the utility value prompted, “Choose a topic from above that is personally useful and meaningful to you (it may be the same topic as before). In 1–2 paragraphs (75–125 words), describe how learning about this topic will be beneficial to you in the future (e.g. education, career, daily life).” Although similar in format, the control prompts simply asked students to identify a topic in class and summarize it in 75–115 words.

**Step One: Identify Logic Models**

Recall that the logic models refer to two types of conceptual path models used in presenting fidelity assessment plans. As such, they lay out the theoretical logic of the intervention. The change model is used to represent what the researcher believes is the causal flow of the intervention by its core components. The operation model is used to operationalize the core components by specifying what can be measured that will represent each component.

**Defining the Change Model**

In deriving the change model from the utility intervention, one core component was identified. The major core component is relevance. As discussed, asking students to consider new material in the context of old material may facilitate increased utility value. Because both quantity and quality of relevance are plausibly important to the development of increased utility value, each was included. Furthermore, because the essays are expected to anchor new material to personal relevance, the number of personal pronouns included in an essay may also indicate relevance. Finally, participants were asked to write a specific amount in each essay. It is possible that students who wrote less in their essays were less likely to produce useful relevance compared to students who wrote more.

Figure 3.4a shows the change model for the summary condition which suggests that summarizing information should lead to typical performance. Figure 3.4b shows the change model for the utility condition and specifies that participation in the utility essay should lead to increased writing about the relevance of psychology to students’ lives, which will then increase their self-reported perceptions of utility value for the course. Ultimately, this will lead to increased interest and course performance.

**Defining the Operational Model**

After defining the core components, subcomponents need to be identified. The core component (relevance) comprises five facets: number of connections, connection specificity, connection personalization, the number of personal pronouns, and number of other pronouns. Together, these facets operationalize the subcomponents of the change model (see Figure 3.5). As a set, the indicators of each
Figure 3.4 Conceptual Logic Model for the Utility Value Intervention. Accurately assessing fidelity requires a strong understanding of the theory of change underlying the intervention. The control group (a) and treatment group (b) may have different theories of change. The models themselves are considered conceptual logic models because they convey the individual pieces of the intervention groups as well as the theorized causal path.

Figure 3.5 Operational Logic Model for the Utility Value Intervention. After settling on a conceptual model, it’s necessary to identify measurable aspects (or sub-components) of the intervention. The operational logic model adds a description of the sub-components of each major point of the intervention. For example, the major core component of the example intervention is “Relevance” which is operationalized through five subcomponents: connection frequency, specificity, personalization, personal pronouns, and other pronouns. In this figure, black-rounded boxes represent the intervention processes, gray-rounded boxes represent psychological processes, and white-rounded boxes represent outcomes.
Incorporating Intervention Fidelity

component give us an idea of the fidelity to that particular component as well as diagnostic information. They provide more detailed information about specific components which can serve to focus attention when modifying the intervention in the future.

**Step Two: Identify Appropriate Fidelity Measures**

After defining the conceptual framework of the intervention, it was necessary to identify and develop measures of the various facets of fidelity. In the focal intervention, a number of measures relied on the qualitative analysis of student essays, which necessitated the development of a coding rubric (see Kosovich, 2013 for the full set of measures).

Of the five relevance writing facets, the first three were measured using the coding rubric. First, the number of relevance connections was counted. A connection was considered to exist when a distinct relationship between course material and unrelated material is discussed. For example, a connection was “Now that we have learned about encoding, short term memory and long term memory, I studied very differently for this exam than I would have in the past.” Second, connection specificity was included as a gauge of response depth and was measured on a seven-point (with half-point increments) ranging from 0 (Essay contained no connections) to 3 (Essay connections are specific—it explicitly states that two or more things—e.g. events, topics—are related and provides a specific example). For example, a specific connection would be:

> What made me the most nervous about being a freshman in college was being able to pick-up good study habits. After learning and reading about how memory works I decided to try and study a few days before our first exam. On the night before our exam I went back and reviewed everything I had been studying the previous days before and it had felt like I could recall everything very well.

(\textit{Intervention participant, Specificity Rating 3/3})

Third, connection personalization was also measured. It was included because connecting material to personal topics may be more useful than connecting to abstract topics. The connection personalization dimension is also measured on a seven-point scale and ranges from 0 (Essay contained no connections) to 3 (Essay contained strong personalization—it provides a specific instance or example of the topic's personal relevance to the person or significant other—e.g. mother, sibling, friend—rather than general for everyone—why the content is important for this person in particular). For example, a personalized connection would be:

> This material is relevant to my life because a large part of it consists of ways to improve memory and retain information. As a freshman, this can be very
important to me because I am adjusting to college classes and how I study the concepts and information from these courses.

(Intervention participant, Personalization Rating 2.5/3)

One important characteristic to note from the set of essays in the current study is that specificity and personalization tended to occur together. Although some essays with high personalization and low specificity did occur, they were uncommon.

The fourth and fifth facets incorporated in the relevance components were drawn directly from essay text. The fourth facet was the number of personal pronouns in the essays. Personal pronouns were all instances of “I,” “Me,” “My,” and “Mine.” The fifth facet was the number of other pronouns in the essays. Other pronouns were defined as all instances of “We,” “Our,” “Us,” and “Ours.”

In addition to the two intervention-related components, an additional pair of measures was included for assessing group similarity. Both of these measures provided an opportunity to measure components that should have been roughly equal between the treatment and control groups. Thus, they served a dual purpose of assessing group differences in fidelity analyses (see Step Four) and as covariates for outcome analyses (see Step Five). First, the average relevance essay word count (as opposed to the goal setting essay word count used above) was used to determine if similar amounts of writing is being done in all conditions. Second, essay writing quality was also included on the coding rubric described earlier. Writing quality was rated on a seven-point scale, half points are used, ranging from 0 (No Response) to 3 (Essay contains groups of sentences that are logically related and clearly understandable). Although writing quality can be represented with a much more detailed set of indicators, this indicator was only meant to provide a measure of drastic differences in writing quality, rather than more subtle nuances.

Step Three: Conduct Psychometric Analyses

Evaluating the Psychometric Properties of the Fidelity Measures

Step Three of the five-step framework specifies that the psychometric properties of any measures used must be examined. Because the primary focus of the current study is fidelity analysis, reliability and validity evidence for the fidelity rubric is of primary importance. Inter-rater reliability, using adjacent percent agreement (i.e. ratings were within one point of each other), was calculated to estimate how consistent raters were at reliably producing ratings of essay elements. As with scale items, determination of reliability is necessary for assessing whether or not raters can dependably produce ratings. Poor reliability results in scores that are not interpretable because the values depend on situational factors rather than intervention factors. Table 3.2 contains descriptive statistics for the various fidelity measures.

Fidelity indices that required reliability estimates include (a) the number of relevance connections, (b) the specificity of those connections, (c) the degree of
Incorporating Intervention Fidelity

personalization of those connections, and (d) writing quality. The other fidelity indices (word count, number of personal pronouns and other pronouns) do not require reliability estimation because they are observed variables and are assumed to have no measurement error; these measures were counted using a computer.

A total of 750 essays were rated during the course of this study. Due to the number of essays, every essay could not be rated by two people. Instead all of the essays were rated by at least one rater, and a series of random samples—totaling 20% of the essays—was drawn from the population of essays and rated by three additional raters. Overall, adjacent percent agreement across raters was acceptable by commonly reported reliability values (Stemler, 2004; Frick & Semmel, 1978) suggesting that different raters could consistently agree about 87% of the time within one point (on a seven point scale) on the fidelity elements.

Step Four: Conduct within- and between-Group Fidelity Analyses

How Well Was the Intervention Implemented Based on Different Indexes of Achieved Relative Strength (ARS)?

Table 3.3 contains basic descriptive statistics for the fidelity measures for each group. To determine how well the intervention was implemented in the study, achieved relative strength was examined in three ways using average, absolute,
and binary complier indices to compare the treatment and control groups. In particular, ARS comparisons were made for: word count, writing quality, number of connections, specificity, and personalization. Table 3.3 contains average, absolute, and binary complier contrasts for all five of the fidelity measures. Figure 3.6 compares the common components (e.g. writing quality and word count) to fidelity components (e.g. connections, specificity, and personalization) in standard deviation units. This comparison demonstrates that group differences on the important intervention fidelity factors are 3 to 5 times larger than differences on the common components.

### Average Fidelity

The first analysis is the average ARS which compares experimental group means to determine differences in fidelity. Previous research utilizing ARS indices suggested using Hedges’ $g$ (Hulleman & Cordray, 2009), which is an effect size measure from the family of standardized mean differences. Hedges’ $g$ is calculated using the pooled standard deviation of the two groups being compared; however, its...
calculation assumes relatively similar standard deviations and distributions (Kline, 2004). Experimental designs are particularly problematic for such assumptions because treatments can cause different distributions and variances resulting in a pooled standard deviation that does not represent either group effectively. To account for such problems, an alternative to $g$, Glass’s $\Delta$, was proposed (Smith & Glass, 1976). Glass’s $\Delta$ uses the standard deviation of the comparison group rather than a pooled standard deviation because it assumes that the typical standard deviation in non-experimental settings is more like that of the comparison group. For outcome effect sizes, $\Delta$ uses the standard deviation from the control group, thus approximating the distribution of a group without any treatment. Kline (2004) suggests reporting $\Delta$ using each of the groups’ standard deviations because the selection of a comparison group can cause drastically different values. For example, in an experiment that has three conditions $\Delta_1$, $\Delta_2$, and $\Delta_3$ would be calculated using the standard deviations of Group 1, Group 2, and Group 3 respectively.

Intervention fidelity is a special case of experimental comparison because it focuses on adhering to conditions that are different from normal. Because fidelity researchers are interested in the relative difference in fidelity between the treatment and control groups, rather than comparing treatment to a baseline condition, the logic of $\Delta$ becomes reversed. That is, the appropriate baseline condition for
understanding adherence to the intervention model is the intervention condition rather than the control condition. Thus, the standard deviation of the intervention group will be used in the calculation of the achieved relative strength effect size resulting in delta-fidelity:

$$\Delta_f = \frac{\bar{M}_{tx} - \bar{M}_k}{s_f}$$

Where $\bar{M}_{tx}$ is the mean of the treatment group with the highest expected fidelity, $\bar{M}_k$ is the mean of any comparison group (either other treatment or control groups), and $s_f$ is the standard deviation of the highest-order treatment condition (the group expected to have the highest fidelity of all treatment conditions). In cases of hierarchically ordered intervention groups, researchers should consider calculating $\Delta_f$ using the standard deviation of the group with the most core components rather than the standard deviation of the control group. The important distinction between Glass’s $\Delta$ and $\Delta_f$ can be made at a conceptual level. The effect size, $\Delta$, was designed to test the departure of a condition from a theoretical baseline. For typical experiments, researchers are testing departure of a treatment group from a control group. For fidelity analyses, however, researchers are testing for infidelity from the intervention. Because the ideal intervention does not exist, $\Delta_f$ treats the highest order condition as the best approximation of the theoretical intervention. Therefore, Glass’s logic was adapted for the fidelity framework to produce $\Delta_f$ (Kosovich, 2013). As a side note, these effect size measures should not be considered with the same treatment-counterfactual logic that governs intervention outcome analyses. ARS contrasts take a step outside of strong causal inference in an effort to describe and explore potential pathways where intervention mechanisms operate. It is worth repeating that the interpretation of $\Delta_f$ differs from Glass’s $\Delta$ because we are now interested in departure from fidelity rather than departure from a control, counterfactual, or baseline.

Table 3.3 contains the several different methods for calculating average ARS contrasts. Of primary interest are the effect sizes contained in the $\Delta_f$ column; the other effect sizes are supplied for comparative purposes (see also Figure 3.6). As expected, the utility condition had drastically higher values on the three major fidelity indices than the control group: connections ($\Delta_f = 3.07$), specificity ($\Delta_f = 4.73$), and personalization ($\Delta_f = 4.29$). The utility condition was also somewhat higher on word count ($\Delta_f = 0.78$) and writing quality ($\Delta_f = 0.16$). Although there was a sizeable difference between the control and treatment groups on word count, it appeared that the difference was an artifact of the intervention prompt rather than intervention effects. During the first dose of the intervention, participants in the control group were asked to write “about 1 or 2 sentences” per topic (4 topics) presented to them whereas participants in the treatment conditions were asked to write one or two paragraphs about a single topic. The differences in prompt instructions appear to be the culprit for word-count differences because
the differences between groups vanished in the second intervention dose which indicated a word range. Unsurprisingly, the utility group ($\Delta f = 6.5$) had drastically higher relevance than the summary group as expected.

**Absolute Fidelity**

As a secondary measure of local fidelity, a subset of absolute fidelity indices were also calculated using a variant of Hedges’ $g$ adjusted for proportions (Hulleman & Cordray, 2009). Absolute fidelity was determined by comparing proportions of students in the treatment and control groups who achieved a maximum value on the particular measure being examined. For example, writing quality, specificity and personalization all had maximum values of three. With indices without an absolute value, an empirical or theoretical absolute was used instead of an actual maximum value. For connection frequency, we chose three connections or higher based on the frequency distribution of responses. For word count, we used the directions given to students as a guide (i.e. 250 words).

**Binary Complier**

Similar to absolute fidelity, binary complier indices require categorization of individuals as compliers or non-compliers. In other words, the researchers or evaluators create a theoretically and/or empirically driven standard for which *minimally acceptable fidelity* has been achieved. This particular type of index allows researchers or evaluators to potentially detect if a critical mass of intervention participation has occurred. In this case, the users must set a cut-off value that identifies individuals as compliers. For example, individuals in the current example were considered writing quality compliers if they achieved a rating of at least 2.5—this was the point in the scale where the writing was expected to be a coherent idea. Similarly, specificity, personalization and number of connections were all assigned a value of two as indicating minimally acceptable fidelity. Both specificity and personalization used two as a cut-off because that is the point in which the individuals clearly articulated an experience, whereas more than a 2 indicated some greater degree of elaboration (which was desirable but not necessary). Similarly, number of connections used a value of two because the activity required at least two connections to be made (again, more was desirable but not necessary). Finally, word count used 150 words as a cut-off because that was the minimum recommendation for the activity in both groups—thus 150 words or more was an indicator that the individual was following directions.

**Step Five: Link Fidelity to Outcomes**

Because most of the subsequent chapters in this volume are examples of linking fidelity measures to outcomes, we only briefly describe this step for our current
example. Using the example data, we restricted our analysis to the intervention group because the control-group variance was essentially zero. In prior studies of the utility value intervention, students’ self-reported perceptions of utility value mediated the effect of treatment assignment on self-reported interest. In Step Five of the fidelity framework, we are interested in testing if variation in treatment uptake can predict differences in outcomes. To this end, we conducted a mediation analysis of the effect of fidelity on student interest as it is transmitted by utility value (Togfighi & MacKinnon, 2011). Using multiple regression, we first tested the effects of our fidelity components on self-reported utility value. This model included all of our fidelity measures, but was restricted to the intervention group. Ultimately there was a positive relationship between rated connection frequency and utility value ($b = .22, p = .03$). We then tested the effects of our fidelity components on self-reported utility value. This model was similar to the first, except that it also included utility value as a predictor. In this case, there was only a significant effect of utility value on interest ($b = .60, p < .001$). We then computed the indirect effect of rated connection frequency on interest by multiplying the two regression coefficients together and calculated confidence intervals for the effect size. This analysis revealed that the effect of fidelity on self-reported interest was mediated by utility value ($b \times b = .13, 95\% CI [0.03, 0.25]$). This analysis provides two important pieces of information. First, that variation in treatment fidelity (i.e. number of connections written) predicts variation in the theoretical mechanism of the intervention (i.e. perceptions of utility value). Second, it shows that this effect is transmitted to our more distal outcome, interest, through its impact on utility value. Thus, for the example presented, we can say confidently that our treatment and control groups were highly differentiated, and that variation in treatment uptake related to improved outcomes.

Conclusions

The intent-to-treat analysis will remain the gold standard for program evaluation, by providing evidence for whether an intervention works. However, the precise measurement of intervention fidelity is critical for evaluating how and under which conditions these interventions work. Analyses of fidelity data complement the strong causal inferences afforded by the intent-to-treat analyses. The five-step framework for assessing intervention fidelity provides evaluators with crucial guidance on how to assess intervention fidelity. To be effective, this framework should be used early in the process of developing intervention evaluations. There are many reasons why well-crafted interventions may or may not work. The rigorous assessment of intervention fidelity is critical for understanding either outcome.

In this chapter we have described the five-step framework for intervention fidelity. Many of the subsequent chapters focus on linking fidelity measures to outcomes, which is step five in our framework. The intent of this chapter was to clarify why steps one through four are critical for the effectiveness of any method
used for this purpose. Without a clear understanding of how the intervention should work, careful consideration of the instruments measuring the core components, an understanding of the psychometric properties of those instruments, and an understanding of the levels of fidelity measured by those instruments both within and between the treatment and control groups, the role of fidelity in understanding outcomes of an intervention are dubious at best.

References


CONTRIBUTORS

Jill Allor  
Southern Methodist University

S. Natasha Beretvas  
The University of Texas

Laurie Bozzi  
Abt Associates, Inc.

Anna Mari Fall  
The University of Texas

Frank M. Gresham  
Louisiana State University

Chris Hulleman  
University of Virginia

Richard Hurtig  
University of Iowa

Jeff Kosovich  
University of Virginia

Carolyn Layzer  
Abt Associates, Inc.

Nancy Scammacca Lewis  
The University of Texas

William Murrah  
University of Virginia

Cristofer Price  
Abt Associates, Inc.

Greg Roberts  
The University of Texas

Arthur Smith  
Abt Associates, Inc.

Lynne Stokes  
Southern Methodist University

Fatih Unlu  
Abt Associates, Inc.

Sharon Vaughn  
The University of Texas

Vivian Wong  
University of Virginia

Keith Zvoch  
University of Oregon