A Procedure for Assessing Intervention Fidelity in Experiments Testing Educational and Behavioral Interventions

Michael C. Nelson, BS
David S. Cordray, PhD
Chris S. Hulleman, PhD
Catherine L. Darrow, PhD
Evan C. Sommer, BS, BA

Abstract

An intervention's effectiveness is judged by whether it produces positive outcomes for participants, with the randomized experiment being the gold standard for determining intervention effects. However, the intervention-as-implemented in an experiment frequently differs from the intervention-as-designed, making it unclear whether unfavorable results are due to an ineffective intervention model or the failure to implement the model fully. It is therefore vital to accurately and systematically assess intervention fidelity and, where possible, incorporate fidelity data in the analysis of outcomes. This paper elaborates a five-step procedure for systematically assessing intervention fidelity in the context of randomized controlled trials (RCTs), describes the advantages of assessing fidelity with this approach, and uses examples to illustrate how this procedure can be applied.

Introduction

Reviews of the literature have shown that higher fidelity of implementation is associated with greater treatment effects.[1–3] Yet, it has also been the case that the assessment of fidelity has been far from universal and there has been great variation in how researchers conceptualize and measure

Address correspondence to Michael C. Nelson, BS, Department of Psychology and Human Development, Vanderbilt University, 201-A Hobbs, Nashville, TN 37203, USA. Phone: +1-330-7058840; Email: nelsonm555@gmail.com.

David S. Cordray, PhD, Department of Psychology and Human Development, Vanderbilt University, Nashville, TN, USA. Email: david.s.cordray@vanderbilt.edu

Chris S. Hulleman, PhD, Center for Assessment and Research Studies, James Madison University, Harrisonburg, VA, USA. Email: hullemcs@jmu.edu

Catherine L. Darrow, PhD, Abt Associates Inc., Cambridge, MA, USA. Email: Catherine_darrow@abtaccoc.com

Evan C. Sommer, BS, BA, Department of Psychology and Human Development, Vanderbilt University, 1900 Richard Jones Road, Nashville, TN, USA. Email: evan.c.sommer@vanderbilt.edu

An earlier version of this paper was presented at the Society for Research on Educational Effectiveness 2010 Conference.

Journal of Behavioral Health Services & Research, 2012. 1–22. DOI 10.1007/s11414-012-9295-x

Published online: 31 August 2012

fidelity of implementation[3–5]. O'Donnell[4] lists seven definitions of implementation fidelity from health literature and six from educational literature, pointing out that the former tended to focus on the presence of program components, while the latter often emphasized changes in quality of practice. In the context of programs implemented in community organizations, Fixsen et al.[6] separate fidelity into two types: personnel fidelity (the implementation of the actual intervention) and organizational fidelity (the implementation of intervention supports such as training and coaching). A frequently referenced and especially comprehensive definition of fidelity[1] distinguishes five types: adherence (did implementers do what was expected?), exposure (did participants receive as much as expected?), quality of delivery (did implementers perform activities in the manner expected?), participant responsiveness (did participants follow through as expected?), and program differentiation (did the treatment condition differ from the control condition as expected?).

The multifaceted nature of fidelity, together with the absence of a unified approach to fidelity within and across research disciplines (e.g., education, health, social programs), have led researchers to employ a variety of terms (e.g., treatment integrity, adherence, competence, compliance) and even more varied operational definitions of fidelity and methods of assessing it. The confusion around fidelity and its measurement can best be addressed through an explicit, systematic framework for assessing fidelity of implementation; without such a framework, researchers may find it difficult to formulate a comprehensive plan for assessing fidelity and to follow that plan consistently. Furthermore, such non-systematic assessment of fidelity of implementation can decrease the quality and usefulness of fidelity data. The goal of this paper is to generate more precise guidelines for assessing fidelity the context of randomized controlled trials (RCTs) of educational and behavioral interventions. Specifically, this paper outlines a process for assessing *intervention* fidelity, which is the extent to which an intervention's core components have been delivered as prescribed and differentiated from the comparison condition.

Systematic Fidelity Assessment and Causal Inference

Although some guidelines for fidelity assessment do exist,[3,4,8–10] their usefulness is limited by their generality: often being developed based on literature from an array of fields and study types, with sufficient breadth that researchers could apply them to an equally broad array of study contexts. The generality in fidelity measures is matched by the tremendous variability in definitions of fidelity, which can lead to confusion in exactly what should be measured as a part of fidelity assessment. Without a focus for fidelity assessment, measuring anything related to implementation is justifiable, thereby leaving the researcher with a haphazard set of measures limited only by data collection resources. To combat these issues, this paper articulates a logical framework and processes for assessing intervention fidelity that strictly parallels the logical framework and processes for conducting randomized control trials (RCTs). While the framework does not answer all questions with respect to how a researcher should go about assessing fidelity in a particular RCT study, it does provide a sequence of steps that researchers could follow to be systematic in that assessment. The guidelines also serve to inform intervention designers as to the types of information researchers will need in order to design studies that systematically assess fidelity to their interventions.

Systematic fidelity assessment follows a pre-specified sequence of steps with a clear beginning point and final objective. A detailed description of the intervention is crucial, both in terms of its structural core components and the underlying theory for its processes. This allows researchers to specify what is meant by fidelity of implementation: the five-step model of fidelity assessment focuses on the core intervention components that then become the appropriate targets of fidelity measurement. This model includes linking the intervention core components to outcomes. Although not the only way to conceptualize and measure fidelity, such a systematic focus enables

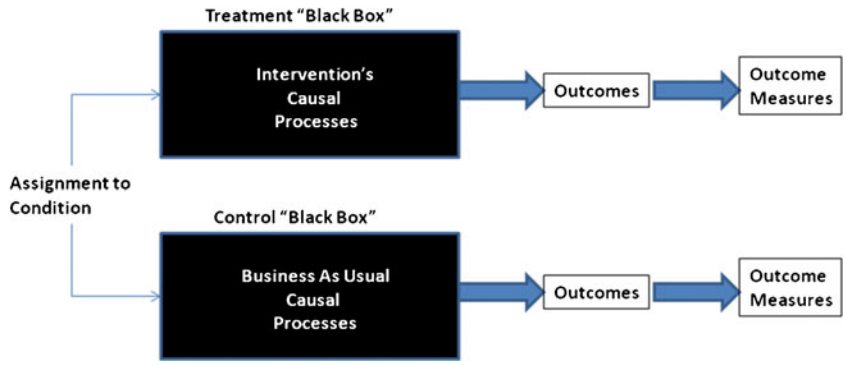
researchers to prioritize and justify what is measured. Importantly, this definition of intervention fidelity is distinct from the assessment of organizational fidelity,[6] which is limited to the implementation of supportive elements like staff selection, administrative training, and the provision of resources.[7] These implementation core components[6] are not part of the intervention model per se, but rather help support the implementation of the intervention core components.

The result of this systematic process is that it bolsters the argument for internal validity by providing explicit evidence about intervention effectiveness as a result of implementation. However, true claims to cause and effect are most strongly made through the randomized experiment, and more specifically the intent-to-treat (ITT) model. As detailed by both Rubin and Holland,[11] random assignment of multiple individuals to mutually exclusive conditions presents the ideal solution, given that randomization diminishes the probability of different characteristics of individuals (or setting or timing or any other element that is randomized) covarying with the cause. Statistical treatment of such a study's results allows for the determination of the average causal effect,[11] that is, the difference between mean outcomes for the group that experienced the cause and the group that did not. Strictly speaking, what differs between the treatment and control groups is not the exposure of its members to the cause but the *intent* to be exposed (i.e., the individual's assignment to condition). Everything that follows assignment to condition may be described as the effect of assignment.

The ITT experimental model is the gold standard for determining the effects of causes.[12] However, by enclosing everything between the assignment to condition and the experimental outcomes within a metaphorical "black box,"[13] there was no conclusion why intervention effects did or did not occur. In Figure 1 (and all others unless specified otherwise), time moves from the left side to the right, and arrows point from causes to their effects. The standard ITT experimentation model does not measure anything inside the black box, so it is not possible to describe the processes, including intervention fidelity, that led to the measured outcomes of the intervention (nor of the control condition). Because implementation necessarily occurs after assignment, intervention fidelity is enclosed in the black box and can be described as an effect of assignment. However, this also means that the effects of variations in fidelity cannot be established using the ITT model, as this variation occurs after assignment to condition. In other words, variations in implementation are likely due to non-random factors such as contextual and personality variables exogenous (or outside of) the ITT framework.

Figure 1

The ITT experimental model



Why Assess Intervention Fidelity?

While not strictly causal analyses, fidelity assessment can supplement the ITT model by filling in the black box with a model of the intended intervention processes, and by measuring these processes to determine the extent to which they were actually present during the experiment. It is by means of such explanations that researchers move from pure statistics to conclusions about why the intervention worked (or not), and the implications for future research and real-world knowledge. Explaining experimental results can also have important influences on policy and funding decisions: explaining non-significant experimental results as simply as "the intervention doesn't work" may result in the loss of support for its funding and implementation in the real world. Fidelity assessment can elevate the inevitable attempts to explain outcomes above subjective opinion. In fact, useful information can be gained from experiments with "failed" interventions by describing where the implementation deviated from theory, if applicable. If an intervention was not implemented as intended, then explaining null results by characterizing the intervention as inherently ineffective may not be appropriate. Furthermore, identifying components that were implemented poorly may guide future research. Not only does measuring fidelity allow one to confirm construct validity within the experiment,[7] it also assists in generalizing the results of experiments with "successful" interventions by describing what exactly was implemented and worked, and thus what should be replicated. Note how this echoes the call from Institute of Education Sciences (IES) to identify "What works in education, for whom, and under what circumstances".[12 (p. 3)]

However, measuring fidelity in the treatment condition, and indexing the degree to which the treatment was implemented as intended, is not enough. In order to make the causal claim that the presence of the intervention causes differences in outcomes, the extent to which intervention core components exist within the control condition also needs to be assessed. As with most educational and behavioral interventions, the core components are unlikely to be completely novel and may overlap with best practices or business as usual. Thus, core components of the intervention may appear to some degree in the control condition, resulting in less contrast between conditions and possibly weaker effects. For example, when studying the effects of an alcoholism treatment program delivered in a particular homeless shelter, even if admission to the shelter is randomly assigned, one must consider both the extent to which shelter residents receive treatment and the extent to which those assigned to the control condition receive similar help from other venues or even mistakenly from the treatment facility. Thus, fidelity assessment *in both the treatment and counterfactual conditions* allows the researcher to capture intervention strength as implemented (i.e., achieved relative strength [ARS]).[13,14]

Defining fidelity in the treatment group relative to the control group is entirely consistent with and extends Rubin's[11] model for assessing effects: just as Rubin's model defines effects as the difference of effects between conditions, it also follows that causes should be defined as the difference of causes between conditions. The researcher can now attribute the way effects differed to the way causes differed, completing the model. This requires valid and reliable measures of core intervention components that are linked to equally valid and reliable outcome measures. Fidelity measures must contain at least some indices that are sufficiently general (i.e., construct-based) to be applied to the control condition as well as the treatment condition. Finally, the extent to which intervention constructs are present in the two conditions must be differenced to calculate the strength of the intervention that was achieved in the treatment condition relative to the control condition. Calculation of intervention ARS is further described as part of the discussion of Step 5.

What Is Intervention Fidelity?

Intervention fidelity is the extent to which an intervention's core components have been implemented (and differentiated from control conditions) as planned. Intervention fidelity fully opens the black box by measuring the processes linking implementation and outcomes. The rationale for considering the entire contents of the black box as *intervention fidelity* is twofold.

First, the initial materials and activities of interventions are selected by their designers based on a theory of processes. If these processes are not present during treatment as anticipated by the designer, this is evidence that the intervention has not been implemented entirely as intended. Second, fidelity assessment is the measurement of the causes of effects, and the initial components of an intervention are rarely thought to directly and immediately cause the outcomes. Instead, the effects of the initial cause proceed through a chain of causes and effects, as with dominoes in a line, until the last cause results in the effect of central interest, the outcomes. This mirrors somewhat the fidelity measurement approach advocated by Bellg and colleagues:[8] the fidelity checklist they developed for general health behavior interventions[15] includes sections assessing not just components like the intervention model, training of providers, and treatment delivery, but also intermediate components of treatment receipt (e.g., subject's comprehension of and ability to implement intervention activities) and treatment enactment (e.g., subject's performance of intervention skills).

To illustrate, an intervention might involve training parents to provide their children with rewards for making healthy food choices (such as eating more fruits and vegetables), with the intended outcome being improved child health. A typical assessment of implementation fidelity might examine how the training was conducted and whether parents offered rewards as intended. An assessment of intervention fidelity could extend the examination to include whether children took advantage of the rewards, the amount and frequency of fruit and vegetable consumption, and even latent constructs such as improved attitudes toward healthy eating (if intended by the designer). Intervention components like these, between initial implementation and outcomes, are often (correctly) considered intermediate outcomes, but they are also intermediate causes: just as the presence of rewards is necessary to cause increased healthy food choices, the presence of increased fruit and vegetable consumption is necessary to improve child health, and the measurement of both of these causes is necessary for the researcher to determine whether the intervention was in place as intended. Likewise, the participants in an intervention often may be considered intermediate implementers, their responses to the initial causes constituting the further implementation of the intervention. These conceptualizations of fidelity are consistent with the inclusion of process-related criteria, like participant responsiveness in fidelity frameworks;[4] arguments for including treatment receipt and enactment in a full implementation model;[16] and with recommendations issued by the National Institutes of Health (NIH) Behavior Change Consortium for best practices.[8]

In practical terms, researchers may face several barriers to fully measuring the contents of the black box, including limited resources and incomplete understanding of the intervention. A logic model of the intervention may serve as a guide here (described in more detail in Step 1), directing the researcher to channel resources toward measuring the most essential components under the best available understanding of intervention processes, while indicating which particular components remain unknown. Another concern is that assessing fidelity inside the black box involves essentially correlational analyses. Statistical methods and software are available for conducting analyses on appropriate data to provide evidence of causality, and researchers doing so are to be applauded. However, this becomes more challenging under the complex conditions of a field study, and it is understandable that such analyses are less likely to be applied to a secondary research question like fidelity. Nonetheless, even in the absence of proof of causality between components, it remains incumbent on researchers to measure the existence of components: if researchers spend pages justifying funding requests and explaining results based on a theory that A causes B and B causes C, they should at least collect evidence that B exists.

The driving theme of this paper is the imperative that researchers be as careful and complete in measuring causes as they are effects, and so intervention fidelity calls for measuring every relevant cause, to the extent possible, even if it is also an effect. Some researchers who do not conceptualize fidelity in this way may object to applying the label "fidelity" to such assessment, even after

distinguishing it as intervention fidelity. So long as researchers agree that measuring causes is essential for understanding what works in the context of an RCT, the particular terminology is less important. Furthermore, the value of the five-step process described herein is not contingent on acceptance of this definition of fidelity. Rather, model-based fidelity assessment depends first on the intervention models, which may be specified with as much or as little elaboration of processes as desired.

The Five-Step Model of Intervention Fidelity Assessment

Fidelity assessment, by looking inside the black box to measure the extent of implementation, can *explain* the effects of causes. Such a model-based approach to fidelity assessment can be more systematic and serve as a template to intervention developers and researchers. To that end, described below is a five-step process for assessing intervention:[17] (1) specify the intervention model, (2) identify appropriate fidelity indices, (3) determine index reliability and validity, (4) combine indices where appropriate, and (5) link fidelity to outcomes where possible. These steps are illustrated using examples from a review of published intervention studies[18] and elsewhere.

Step 1: Specifying the intervention model

The precise contents of the black box will depend on the particular intervention because different interventions achieve their effects in different ways. In general terms, when looking inside the black box, one finds the independent variables whose manipulation constitutes the intervention and any mediating variables that ultimately convey the effects of the cause to affect outcomes (i.e., the dependent variable(s) of the experiment).

The contents of the black box, as well as their outcomes, can be represented in two ways. If these elements are represented in practical terms, describing the activities and resources that will be involved in the intervention's implementation within the context of a specific experiment, this is the logic model (sometimes called "operational logic model" or "operational model")[19]. Intervention developers or evaluators frequently create logic models to guide implementation. When the model depicts only the essential intervention elements, representing them in conceptual terms by describing the constructs that underlie activities and resources and effect outcomes, this is the change model (sometimes called "theory of change," "conceptual logic model," or "conceptual model").[19,20] For complete assessment of intervention fidelity, it is necessary to describe both models, beginning with the change model.

Specifying the change model

The change model consists of whatever constructs the intervention designer believes will be involved in the causal process, in whatever causal relationships the designer envisions. Most interventions have a change model, though this may only be implicit in the developer's understanding of the mechanisms by which the intervention operates. As noted above, the change model is a hypothetical set of constructs and their relationships created by the researcher; and like any hypothesis, it need not have a rational or empirical foundation nor even be plausible in order to be tested. Realistically, however, the functioning of most interventions does have some rational, theoretical, and/or empirical foundation that can and should guide change model development.

The value of a change model for evaluating social programs has been acknowledged by many in the evaluation field for some time.[20–22] These sources also note the difficultly of developing accurate change models. Despite any challenges that may be encountered, any description of an intervention and its effects should begin with specification of the change model for four reasons:

(1). *Constructs are abstract representations of important processes, which may be generalizable*. If the goal of an experiment is to prove that an intervention "works" in general, not

just that it worked in the particular context of the experiment at hand, the intervention model must be described in general and abstract terms before operationalizing it in the specific and concrete terms used to test it.

(2). *Change models represent a network of causal connections*. The real cause of an effect is a matter of perspective: a pot of water on the stove boils because the stove is turned on , but it also boils because of the fire beneath the pot, because water molecules are excited by the heat from the fire, and because of the laws of thermodynamics as they apply to those molecules. Given the ambiguity of real cause, elements in a logic model are related not by causes but by sequence. In a change model, the designer does not have to define real causes but only indicate their presence conceptually, at whatever level of specificity is desired.

(3). *The change model delimits what is to be measured, when, and how*. Except in the most constrained laboratory conditions, one generally cannot measure every variable that may be present in the real experimental context to determine if it played a role in the causal process. Instead, one attempts to measure only elements that could plausibly play such a role (the "core intervention components"), as envisioned by the designer of the intervention or others testing it. Given that the change model represents all such elements, it can serve as a guide when creating and applying measures.

(4). *The change model can be used to guide analysis of measures*. The causal processes of an intervention might involve numerous variables, and each variable might be measured using multiple indicators, altogether yielding several separate descriptions of the causes of outcomes. By linking all indicators to constructs, the change model can suggest how multiple indicators can be combined to describe a single construct. Similarly, because the change model describes causal links among constructs, it can suggest how to combine measures of separate constructs within the intervention to quantify the cumulative "cause" of the outcome.

In short, the change model differs from and enhances the more widely employed logic model in that it consists of constructs representing generalizable processes rather than specific activities and resources, specifies causal connections rather than mere sequences of events, and includes only core components rather than ancillary supports that can then be used to identify what, when, and how to measure fidelity, and how to link it to outcomes. Specifying this model is crucial in identifying the core intervention components which are theorized to drive the effect of the intervention, and are thereby the focus of fidelity assessment.

Components of a change model

The general structure of intervention change models should be recognizable to those familiar with causal models.[23] At least two types of constructs will be found in any change model: intervention components and outcomes. The intervention components are the change representations of the resources and activities that constitute the "causes" of the intervention, while the outcomes are the psychological and behavioral constructs upon which those components are expected to have their ultimate effect. Another type of change model construct is mediators, that is, any constructs in the causal path between intervention components and outcomes. If a given change model includes mediators, the intervention components cause effects on mediators, which may then cause effects on other mediators until the "chain reaction" of the intervention culminates in the outcomes.*

*It is also possible that an intervention's developer would specify as part of its change model one or more moderators, constructs thought to influence the nature (strength) of the causal relationship between two or more constructs. However, we have omitted discussion of moderators because they are exogenous to the intervention as designed.

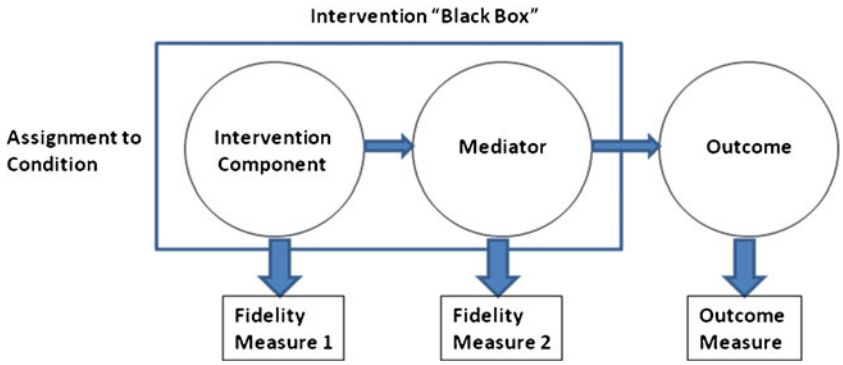
Creating a change model necessitates identifying and describing the intervention's core components, those elements that are unique to the intervention (as compared to a counterfactual condition) and are essential to achieving its effects. Even outside of the five-step procedure, identification and implementation of core components has been recognized as important for achieving intended effects and successfully replicating interventions.[6,10] These components will vary depending on the particular intervention and even on different theories of the same intervention, but the process of identification should begin with an understanding of the theoretical basis for the intervention's form and function (i.e., the change model). Ideally, intervention designers and researchers will develop this understanding collaboratively, discussing each individual's interpretations of the intervention's components and building a consensus of what the key components are and how they relate to one another.[7] Empirical evidence is also valuable: if the intervention is a replication or involves previously studied processes, it may be possible to use implementation and outcome data from earlier research to help determine which components are essential and which are flexible or superfluous.[6] In fact, assessing intervention fidelity through the five-step process facilitates the evaluation of individual core components by linking fidelity data to each component, which in turn allows for the refinement of the intervention model.

Change models may be represented in graphical format, as shown in Figure 2. The black box is not generally represented in the change model diagram, but it is made explicit here to emphasize the transparency of the change model. The number of mediators may be more or less; the number of distinct intervention components and outcomes may be more (but not less). Depicting the change model in this manner is not a mere convenience for the purposes of this paper: It is strongly recommended that researchers develop a graphical representation of any intervention model, supplemented with text description of the model, in order to ensure clarity and mutual understanding of the model among all involved in its development and use.[19] In reviewing articles from the literature,[18] nearly all interventions were with some sort of theoretical justification in narrative form. Yet, as trained readers work independently to derive a precise representation of the change model based on the same narrative, they could produce models with striking and significant differences. Such divergent interpretations among program developers, researchers, and implementers could be disastrous, especially if the incongruity does not come to light until after implementation.

The intervention components and outcomes, respectively, are the first and last links in a sort of chain, called a causal path, that can be connected directly or via other links that are mediators. The above template of change models represents one of the simpler cases in that it involves only a single chain; however, depending on the intervention, the black box could conceal multiple causal paths. An intervention with a single component might have effects on several outcomes via

Figure 2

Template for graphical representation of an intervention change model



diverging causal paths, and a multicomponent intervention's effects might collectively result in a single outcome via converging causal paths. Or each component in a multicomponent intervention might independently affect separate outcomes via parallel paths. Of course, as more complex interventions are considered and additional intervention components and mediators are introduced, a growing variety of combinations of causal paths are possible.

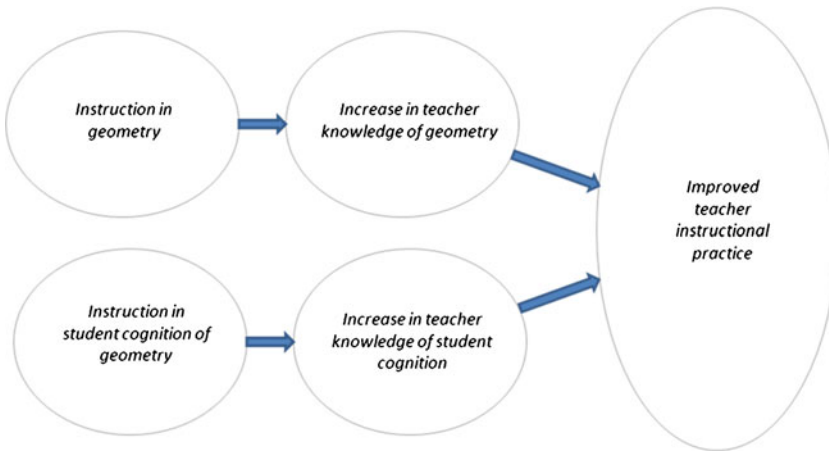
As an example of a change model, the diagram shown in Figure 3 has been adapted from the general model presented for Project LINCS,[24] an intervention that seeks to provide teachers with training that enhances their instruction in the subject of geometry. It is apparent from the change model in Figure 3 that the intervention components (far left) are instruction for teachers in geometry and students' cognitive development in geometry, the impact of which is mediated, respectively, by increased teacher knowledge of geometry and student cognition, to impact the outcome of improved teacher practices.

Had the focus of this study been to study changes in student performance resulting from the Project LINCS trainings, the additional construct of improved student knowledge could have been added at the far left as the outcome, with improved teacher practices being enclosed in the black box as a mediator from increased teacher knowledge constructs to improved student performance. Furthermore, depending on how the developer or researcher conceives of the intervention's processes, the construct of improved student knowledge might be inserted as a new mediator between teacher practices and student performance. The fact that this revision of the model could be continued almost indefinitely (perhaps adding constructs for teacher enthusiasm or student engagement) underscores the fact that the intervention model is almost never obvious and should be made explicit before planning the implementation and measurement.

Specifying the logic model

Once the change model is elaborated, one can operationalize its components for the purposes of a particular implementation. As previously defined, the logic model of the intervention consists of the resources and activities (both of implementers and of participants) necessary to operationalize the change model components for the treatment condition of the experiment. Because the logic model is not conceptual, the arrows in the diagram indicate how components are related in the implementation process over time rather than strict cause-and-effect relationships.

Figure 3
Project LINCS change model



In essence, the logic model serves as a roadmap for implementing the intervention, and it is therefore far more common for a developer to explicitly describe an intervention's logic model. As represented in a logic model, the direct activities and resources that constitute the intervention (inputs like training) lead to activities by participants (outputs like specific changes in teachers' instruction or students' completion of specific assignments) and ultimately to the intended results (outcomes like improved test performance). Yet, in order for the intervention model to serve as the foundation for fidelity assessment, the logic model cannot be independent of the change model. The operational model should be carefully aligned with and even proceed from the change model, each change model component being represented by one or more operational model components.

Figure 4 illustrates how the Project LINCS change model[24] was modified to a logic model (as adapted from the researchers' "applied" model). To explain the diagram, boxes with bold labels represent the concrete logic model components, while ovals with italic labels represent change model components (constructs). Logic model components operationalize aspects of particular constructs. Here, instruction in student cognitive development in geometry is operationalized by the logic model component of a research seminar based on the van Hiele model for such development, while the outcome construct of improved instructional practices is operationalized by three logic model components (what is taught, how it is taught, and characteristics displayed by teachers as they teach). Because causation takes place among constructs, the arrows in the logic model more signify sequence than causation. Note that the changes in teacher knowledge of geometry and student cognition remain unoperationalized constructs in this logic model. This model is simplified; a more realistic logic model describes activities and resources in detail, breaking them down into specific, measurable activities (see facets and indicators under Step 2), a sort of road map to implementation. A more elaborate example of constructs operationalized as logic model components and then measures is given in Figure 5 in the discussion of Step 2.

Both logic and change model components, being opposite sides of the same coin, are contained in the black box (with the exception of outcomes). Consider, too, that one can also develop change and logic models for the expected activities in the control condition; however, rather than creating indices unique to the control condition measuring deviations from business as usual, it is advised that the same general measures of treatment components be applied to the control condition. The difference between the two conditions can then be calculated as the ARS of the intervention (see Step 5 below).

Figure 4

Project LINCS logic model as elaborated from the change model

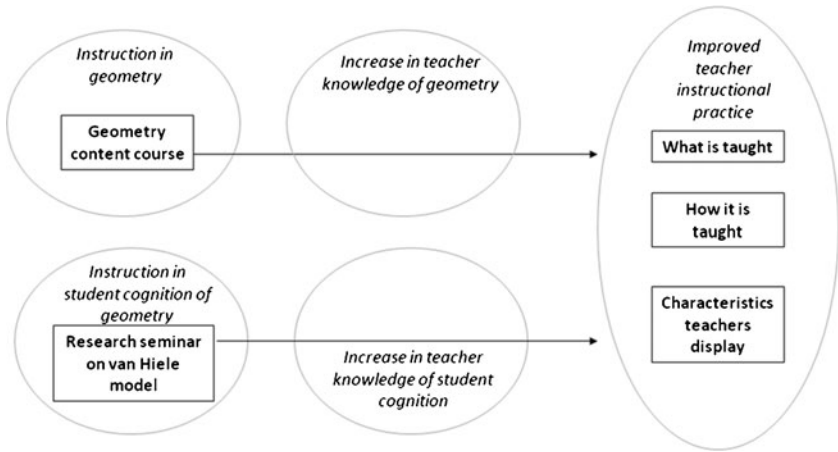
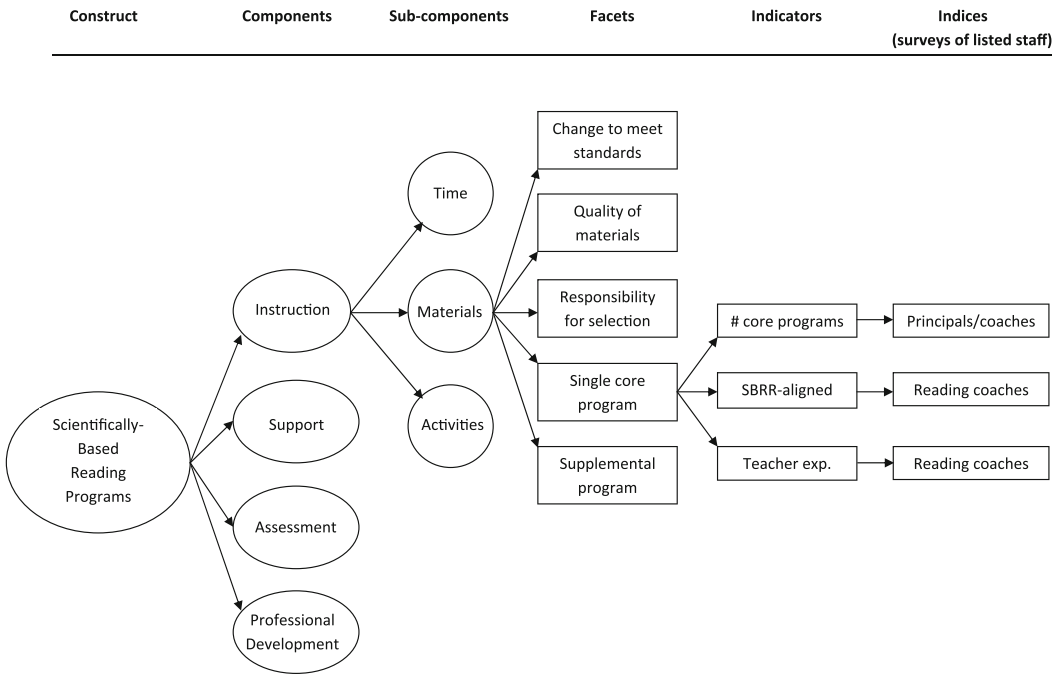


Figure 5

Indices derived from a complex construct in the Reading First Evaluation change model



Once the contents of the black box have been elucidated, they must next be measured so that they can be used to explain experimental outcomes.

Step 2: identifying appropriate fidelity indices

Together, the change and logic models completely characterize the intervention and constitute *the* intervention model*; likewise, the control condition is completely characterized by its own logic models. For the purposes of conducting the experiment, however, it remains to design indices for the measurement of intervention model components. Examples of fidelity indices include self-report surveys, interviews, participant logs, observations, and the examination of permanent products[25] created by the intervention activities. Indices are typically applied to components of the logic model to determine whether they were executed as expected, but such indices can also be conceived of as also indirectly measuring components of the change model. Latent constructs of the change model, such as knowledge or motivation, can also be assessed for expected change by assessing individuals at the appropriate times (as indicated by the change model).

Measures applied to intervention outcomes (outcome measures) produce data that indicate the effects of being assigned to the treatment condition when compared to similar measures of the control. If this analysis indicates that there is a significant difference in outcomes between conditions, then the intervention (as implemented in the particular experiment) is said to have had a

*Note that most models also involve assumptions (e.g., student characteristics) that may not be included in the graphic representation but that should be elaborated narratively.

significant effect. As stated, the ITT model accomplishes this very well and no experimenter would argue with the need for outcome measures.

Yet outcome measures are not the only type of measures that should be applied within the experiment. Measures applied to causes in the model (e.g., fidelity measures) produce data that describe intervention fidelity, the extent to which the intervention as implemented is similar to the intervention as represented by the intervention model. Note that the initial intervention components may not be the only causes in an intervention model: when an intervention component is mediated, the mediator is both an effect of the component and a cause of a further effect. By describing the cause part of the cause-and-effect relationships represented in the intervention model and operationalized in the experiment, intervention fidelity assessment indicates how the outcomes were actually achieved and thus explains intervention effects.

Once the intervention logic models have been explicated, they serve as guides for determining what to measure and how to measure it. It might seem that developing fidelity indices would be a straightforward matter of aligning measures to elements of the logic model, merely quantifying the behaviors, resources, and events listed therein. Yet, the logic model alone is a poor guide: as a practical operationalization of the intervention, it usually includes elements that are not unique to the intervention or are so basic as to not relate directly to core components. Thus, identification of indices must begin with the constructs underlying the intervention as identified in the change model, so that each indicator is associated with core components of the intervention.

The process of deriving indicators from the intervention change model involves differentiating each *intervention component* into any *subcomponents*, deriving from each subcomponent specific *facets*, and finally identifying *indicators* of each facet. Components are simply the major constructs represented in the change model; components are often multidimensional, that is, they consist of relatively independent types of activities. For example, instruction might involve direct lecturing of students as well as designing and coordinating learning activities. These narrower and more homogeneous groupings of related activities within a component are the subcomponents, and they serve as a bridge between the broad constructs in the change model and the practical details of implementation in the logic model. These are facets (not to be confused with the facets of generalizability theory), the specific behaviors, events or resources that constitute the implementation of a subcomponent. Finally, for each facet, the researcher must decide on one or more indicators, which are to be directly measured by indices and constitute evidence of the degree to which each facet is implemented.

The value of this process is that it can ensure that each core component of the intervention is fully assessed, and that any extraneous activities are excluded from fidelity assessment. Assuming that the intervention logic models are complete and consistent, any indicator that cannot be tied back to a core component through this path (indicator to facet to subcomponent to component) can be considered superfluous and should be eliminated; any component with subcomponents that are not operationalized with facets or are not assessed with matching indicators can be considered underevaluated and should lead to the identification of additional indicators.

The Reading First Implementation Evaluation[26] provides a comprehensive example of deconstructing components into indicators. Reading First studied elementary schools as they implemented a range of elementary reading programs sharing several scientifically based components: (1) reading instruction, (2) support for struggling readers, (3) assessment, and (4) professional development. Each of these components was described in terms of several subcomponents, such as reading instruction consisting of (1.1) instructional time, (1.2) instructional materials, and (1.3) instructional activities and strategies. Each of these subcomponents was further broken down into the major measureable characteristics of the subcomponents, with instructional materials being operationalized as five facets: (1.2.1) changing materials to meet standards, (1.2.2) the quality of materials used, (1.2.3) assigning responsibility for selection of materials to appropriate staff, (1.2.4) using a single core reading program across the school, and (1.2.5) using supplemental reading programs for additional, targeted instruction as needed. Each

facet was matched with one or more indicators for measurement, as with the facet of a single core reading program, which had three indicators (with noted indices): (1.2.4.1) the number of core reading programs (items on surveys of principles and reading coaches), (1.2.4.2) whether the core reading program was aligned with SBRR standards (items on survey of reading coaches), and (1.2.4.3) whether teachers were experienced with the core reading program (items on survey of reading coaches). This example is illustrated in Figure 5.

Table 1 shows that the number of indicators per facet varies widely among facets (ranging from 1 to 8). The number of indicators should reflect the complexity of the intervention components, subcomponents, and facets, but a certain minimum number of indicators must be used to assess any one facet in order for measures to be reliable (see Step 3). Also, the total number of indicators across all components is rather large (173). This is a reflection of the great breadth of the Reading First interventions,[26] but also emphasizes the need to follow a systematic process for eliminating unnecessary items.

Step 3: determine index reliability and validity

It is vital that researchers employ fidelity indices that are established to be reliable and valid. In reviewing elementary mathematics intervention studies,[18] manuscripts frequently do not report or reference instrument reliability, and descriptions of index validity are even more rare. Reliability should be established at a minimum, because unreliable measures cannot be valid.[27]

Reliability

Reliability may be enhanced by using multiple methods for measuring individual model components.[28] For example, teacher self-report may be a more biased measure of teacher practices

Table 1
Developing indicators from Reading First components

Components	Subcomponents	Facets	Indicators (average number of indicators per facet)
Reading Instruction	Instructional Time	2	2 (1)
	Instructional Materials	5	15 (3)
	Instructional Activities/ Strategies	8	28 (3.5)
Support for Struggling Readers (SR)	Intervention Services	3	12 (4)
	Supports for Struggling Readers	2	16 (8)
	Supports for ELL/ SPED	2	5 (2.5)
Assessment	Selection/ Interpretation	5	12 (2.4)
	Types of Assessment	3	9 (3)
	Use by Teachers	1	7 (7)
Professional development	Improved Reading Instruction	11	67 (6.1)
Totals: 4	10	42	173 (4.1)

in a certain case, but it may also allow the researcher to get at elements of the implementation that classroom observations may not be able to detect reliably or might even influence inadvertently. In such a case, using multiple measures (or a single measure with multiple items) with minimally correlated measurement error better allows for measurement of the underlying construct. Depending on the indicator and context, multiple measures may be completely different methods (e.g., observations and self-report surveys), similar methods applied to different individuals (e.g., teacher surveys and student surveys), or even merely different items in the same scale.

The number of measures required to achieve a desired level of reliability for measuring a construct may be determined using the internal consistency of the items. The classic approach is to employ Cronbach's alpha.[29] Given modestly correlated items, only a few items can yield a reasonably large alpha coefficient,[30] as shown in Figure 6.

One may also determine the number of required measures or individuals to be sampled for reliably measuring a construct by employing generalizability theory,[31,32] which determines the number of items, observations, or individuals needed to achieve the desired level of generalizability.[33,34]

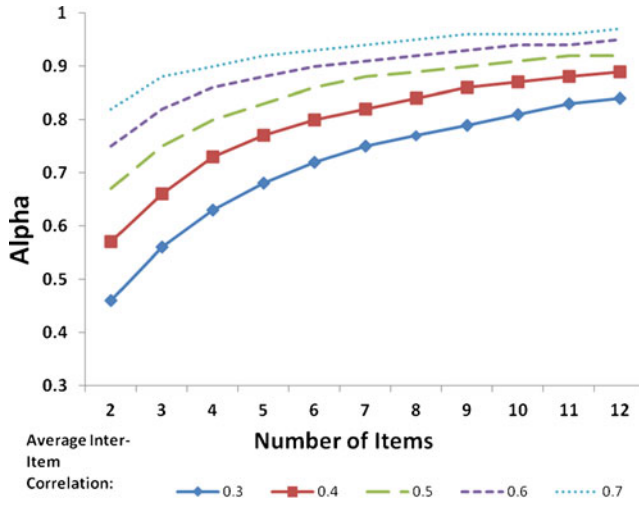
Another issue relating to the number of items used to measure an indicator is the need to fully capture variance in implementation. A single item or single instance of measurement may not fully reflect even a simple activity or may underestimate the variability of its implementation. This deficiency can result in overestimating or underestimating fidelity, and also limits linking fidelity levels to outcomes (see Step 5). The use of summated scales[35] is therefore recommended rather than single items to characterize facets of the intervention. In addition, increasing the number of measurement occasions can increase precision of fidelity estimates, as well as the statistical power to detect relationships between fidelity and outcomes.

Validity

Validating a measure of implementation can be difficult, but it is important that researchers at least attempt to establish content validity, the extent to which a measure adequately represents the construct of interest.[33,36,37] Given that a major purpose of assessing intervention fidelity is to ensure that valid conclusions are drawn from an experiment about the intended intervention, a fidelity instrument lacking in content fidelity would undermine this purpose completely. Just as it is

Figure 6

Determining the number of measures needed to achieve high reliability



valuable for developers to be involved in defining core components of the intervention and creating the change model, developers can make important contributions to the design of valid measures of model components.[28] Indeed, those who created the intervention are likely to have a deep understanding of the constructs' salient, observable features.

Content validity is often assessed during development of the instrument through a validation study, with a panel of experts following a predetermined framework for evaluation[33]. For example, to establish the content validity of a survey of student motivation, experts on the various theories of motivation and its measurement would be employed. Face validity,[38] the extent to which an instrument appears to measure what it purports to measure, may be established for the measure in addition to content validity or instead if limited resources prevent developers from assessing content validity. Face validity can be established by evaluation of the measure by those familiar to the setting or subject to the measure, for example having teachers review a measure of student motivation.

The fidelity measure development and validation process described by McGrew and colleagues[39] was quite robust. Seeking to develop a fidelity measure for the assertive community treatment (ACT) mental health program model, the authors culled important model characteristics from sources in the literature and a one state's quality assurance form. The resulting list of 73 items in seven categories was presented to a group of twenty ACT experts who rated each characteristic's relative importance on a 7-point scale, specified ideal standards for each model characteristic (e.g., size of treatment team, frequency of contacts), and suggested additional characteristics. The IFACT fidelity measure, consisting of three subscales totaling 14 items, was then created based on the ratings, although the feasibility of collecting relevant program data resulted in the omission of many items rated as important. Items were scored from 0 to 1, representing the proportion a program component met the ideal standards specified by experts. Finally, the IFACT was validated by using it to evaluate four generations of ACT programs, each representing a later wave of model rollout. As expected, the results indicated that fidelity as measured by the IFACT was significantly correlated with program outcomes for reducing hospital stays, and both fidelity levels and program impact showed a significant downward linear trend over successive generations.

When developing a fidelity measure specifically for use in an RCT of an intervention, one often does not have the luxury of employing such a thorough development and validation process in advance (for a valuable discussion of issues that may arise and numerous examples of establishing reliability and validity for fidelity indices, see Mowbray et al.[40]). However, following the five-step process helps to ensure valid instrumentation because it requires one to base items on core components identified by program change and logic models (which should already be aligned with models in the literature and expert opinion).

Step 4: combining indices

A thorough assessment of intervention fidelity is likely to involve multiple fidelity measures. This is true even when only a single intervention is implemented once and described by only one change model, because most interventions are designed to include multiple core components and, crucially, most of the constructs that interventions seek to manipulate are complex. Elaborating on the Project LINCS,[24] one may measure separately the presence of, or change in separate constructs in, the change model (geometry instruction and student cognition instruction), which would lead to multiple fidelity scores (an index of geometry course implementation and index of research seminar implementation) that could be considered separately or combined for an overall fidelity score (an index of Project LINCS implementation). Applying multiple measures to multiple components, in addition to bolstering validity, also facilitates fidelity data analysis (Step 5 below): here, one could correlate the overall fidelity score with outcomes to see how much they covary, or one could regress outcomes on the measured components to determine the proportion of variance in outcomes explained by each. The separate analytical methods answer distinct questions about fidelity with important applications.

The example of how Reading First[20] deconstructed a very complex construct into individual facets and indices also illustrates why multiple measures may be required. The construct of "scientifically based reading programs" was conceptualized as being so complicated that it was divided into a hierarchy of four components, ten subcomponents, and 173 indicators. Each indicator was measured by scoring survey items, and these scores could be aggregated (averaged) up to the level of subcomponents (ten scores), components (four scores), or construct (one score). Different score levels could be analyzed with different weighting schemes and linked to outcomes using methods of varying complexity, ranging from simple correlation and regression to an ambitious structural equation model.

Multiple measures may also arise from using multiple methods (e.g., surveys and observations) for the same indicator or the same method with different types of respondents (e.g., surveys of both principals and coaches to determine number of core programs), with the combined measure being more reliable than each taken separately. Crucially, indices of fidelity need to be combined to reflect that they measure the same core component, not merely because they happened to be measured using the same instrument or method. For example, if a child behavioral intervention has two core components for changing behavior toward adults and behavior toward peers, and both components were measured with a child self-report and an observation, then the index of fidelity to peer-related activities will be made up of both the relevant self-report items and the relevant observation items, and the same for adult-related activities. This allows the researcher to evaluate the implementation and effectiveness of each core component individually, and avoids bias due to the type of measure.[7] As Abry et al.[41] demonstrated, creating fidelity indices based on core components instead of by measure increased the variance explained when predicting outcomes in an elementary school intervention.

There are several options for combining indices. Frequently, researchers take a basic approach like summing across indices and calculating the percent of components implemented, as with items in an instrument. Another fairly direct approach is to average the results of several measures, as with multiple observations. Each of these methods of combining indices may be reasonable, yet researchers should be careful that choice of method is driven by the underlying theory of the intervention, or else they may risk giving inappropriate weight to minor indicators. For example, Fuchs and colleagues[42] had observers assess fidelity to their Peer-Assisted Learning Strategies (PALS) intervention using a behavior checklist, with fidelity being calculated as the percentage of checklist items observed. Yet, items on the checklist are given equal weight regardless of the apparent magnitude of their respective roles in affecting intervention outcomes: "teachers provide appropriate corrective feedback," is weighted the same as whether students "replace materials in folder" at activity completion. Both items then receive equal weight when the percentage of present components is calculated, despite the likelihood that teacher feedback has more impact on PALS learning than does putting away materials. As a result, such a checklist may overestimate *relevant* fidelity if superficial, easy-to-implement components are being implemented while more critical, difficult-to-implement components are less reliably implemented.*

Likewise, simply averaging fidelity scores across multiple measures of the same component or different components over time may be inappropriate if the change model indicates that the activities at one time are more crucial to effecting outcomes than the activities at another time. For example, the MAP intervention[43] calls for teachers to be in attendance at four training sessions across 1 year in order to provide them with the skills and knowledge necessary to implement the differentiated instruction program. One approach would be to weight the indicator of teacher attendance equally (0.25) for all four sessions. Yet, the sessions were not equal in terms of content and timing, as displayed in Table 2.

*While this example illustrates the problem in principle, it is unlikely to have inflated fidelity in this particular study given that the proportion of non-core items was relatively small and significant results were obtained.

Note that the initial session is mainly an introductory session on using materials (which would be reinforced throughout the year), while the fourth session takes place after outcome data have already been collected. In discussions with the MAP developers, the researchers were able to determine a weighting scheme (far right column) that more accurately represented the extent that each logic model component (individual sessions) represented the core component (teacher training) as it impacted other constructs along the causal path of the intervention's change model (e.g., change in teacher knowledge, change in teacher behavior).

When different indices are weighted for combination, guidance for determining the weights may come from one or more sources. The intervention developer may be able to explicitly indicate the relative importance of different components, as above, or this information may be apparent from theory. The magnitude of the role individual components play in achieving effects might also be determined empirically by examining the results of previous studies. Another approach would be to conduct a sensitivity analysis using the data from the present experiment, adjusting the weights of components to optimize their prediction of outcomes.

Regardless of the approach for determining weights, as well as the rationale for selecting that approach, they should be reported along with the combined indices for interpretation.

Step 5: linking fidelity measures to outcome measures

Indexing levels of intervention fidelity provides a means for describing the extent to which the intervention as implemented (the actual cause) resembles the intervention as designed (the theoretical cause), which in turn provides an explanation of how assignment to condition resulted in some difference (or lack of difference) between conditions. When uniformly high fidelity is achieved, it is not possible to proceed further: fidelity indices are affirming that the model was followed and accurately describes the intervention process. For example, mean fidelity in the Fuchs et al.[42] study described above was greater than 95% at all observations for both teachers and students, leaving no option but to declare fidelity as high and the model a valid explanation for effects.

Affirmation of total fidelity is a valuable finding, but imperfect fidelity is common, and achieving perfect implementation is increasingly difficult as interventions become more complex. For example, Wilson et al.[44] found that implementation problems were reported 37% of the time for treatment-control comparisons of anti-bullying programs, and Tobler[45] found a rate of 29% among drug prevention studies. A meta-analysis of 59 prevention and health interventions for youth[3] found that most studies achieved no more than 80% implementation and none achieved 100% implementation, although many of these studies were conducted by nonresearchers. There are good reasons to believe that actual rates of flawed implementation tend to be greater than what has been reported: measures of fidelity may be superficial (or at least not checking all components of the intervention, including processes), problems may go unmentioned in cursory reports, and most importantly, variation in fidelity can only be identified when it is measured. It is reasonable to

Table 2
Weighting of training sessions for the MAP intervention

Session	Month	Content	Initial Weight	Adjusted Weight
Session 1	September	Administration	0.25	0.10
Session 2	October	Data use	0.25	0.30
Session 3	November	Differentiated Instruction	0.25	0.50
Session 4	May	Growth and planning	0.25	0.10

expect that fidelity levels may be lower for studies in which they have not even been measured, and meta-analyses have found that treatment effect sizes are substantially greater when implementation monitoring was reported, regardless of actual levels of fidelity[46,47].

The presence of infidelity allows for further analyses linking fidelity measures to outcome measures, revealing how differences in fidelity may be associated with differences in outcomes. Recall that, when intervention models were discussed above, it was noted that it is possible to construct models of both the intervention condition and the control condition based on distinct expectations for each. In principle, then, it is possible to measure fidelity to each of the respective models using different indices; that is, fidelity to the treatment model in the treatment group and fidelity to the control model in the control group. However, an approach more useful for the analysis of fidelity is to apply the indices for treatment fidelity to both conditions. Underlying this approach is the notion that, just as the effect of the intervention is defined as the difference in outcomes between treatment and control, fidelity of implementation may be defined as the difference between components implemented for treatment and control groups. It therefore follows that, when the same or similar indices are applied to both conditions, the difference between the two measures will represent the extent of differentiation between conditions with respect to the intervention model, which is the intervention's ARS.[14] Infidelity within the treatment group will degrade the ARS of the intervention as implemented, as will contamination within the control group.

If this approach is to be employed, the change model will guide the researcher in distinguishing core components of the intervention, which are essential for the intervention's effectiveness and therefore should be measured, from components that are merely supportive or that ought to be common to both conditions and therefore would not contribute to ARS. A caveat to this approach is that some intervention components (or the indices thereof) are too specific for it to be sensible to apply identical indices to both treatment and control. In such cases, it may be necessary to adapt individual items in a treatment fidelity measure for the control, so that adapted items measure not the particular logic model component but its equivalent operationalization of the construct in the control condition. For example, if one is measuring implementation of a particular method of differentiated instruction in the treatment group, the equivalent control index may need to account for more generalized instances of the construct (e.g., ability grouping or individualized assignments). A complete explanation of ARS calculation, together with an example of its application for analyzing experimental findings, is given by Hulleman and Cordray.[14]

It must be emphasized that these analyses are correlational and not causal, but they can provide several tools for interpreting experimental results. When an intervention did not achieve significant results in comparison to control, linking fidelity to outcomes may indicate weak links in the implementation: components that were not fully implemented and thus lessened the difference between conditions. Developers can use this information, together with other evaluation data, to design and incorporate support mechanisms that will bolster implementation of these components in the future. (i.e., implementation drivers).[6]

Conversely, if significant results are achieved despite specific components being weak links in implementation, this is an indication that either: (a) there is potential for boosting effects even further by better supporting implementation of these components, or (b) the components of the change model with lower implementation are not truly core components that are crucial for achieving outcomes. The choice of interpretation between these two options, based on evaluation data, theory, and judgment, will determine whether it is the implementation or the model that is altered.

A third advantage to linking fidelity with outcomes is the ability to determine empirically a cutoff point for sufficient fidelity: that level of fidelity above which variation in implementation has little impact on outcomes. Each component of the intervention may have a different cutoff point, with the most essential components expected to require the highest levels of fidelity. However, evidence that some components have more flexibility could serve as a guide for future

implementers, for example by telling teachers which aspects of a program they can adapt and to what extent.

As noted under Step 3, linking fidelity to outcomes requires that indices have the ability to detect variance in implementation of the intervention. If indices of intervention components involve rating intervention activities on a binary scale (e.g., enacted or not enacted) or with a single item, then there is a risk that the index cannot detect sufficient variance in implementation to link it to outcomes. It is therefore wise to employ counts, Likert-scale items[48], and summated scales[35] to ensure that all relevant variability is captured and can be linked to outcomes.

Examples of researchers statistically linking fidelity measures to outcomes are not abundant in the education research literature, with most researchers merely noting potential weaknesses in implementation and speculating on possible impacts. Among these examples is a randomized field trial by Connor et al.[49] that sought to establish the effectiveness of Assessment to Instruction (A2i) web-based software to guide teachers in properly differentiating early reading instruction. A2i algorithms are based on prior research that students with low letter–word reading skills benefit more from instruction that is teacher-managed (TM) and code-focused (CF), while more advanced readers benefit from increased amounts of child-managed (CM), meaning-focused (MF) instruction. A2i assesses students regularly and provides teachers with daily child-specific recommendations for the amount of each type of instruction needed, optimal groupings of students for reading instruction, and recommended lessons and activities drawn from the school's reading curriculum. Training consisted of two workshops, biweekly coaching from researchers, and collaborative professional development, all delivered to treatment but not control teachers.

Fidelity of A2i use was measured as minutes; teachers were logged in to the computer system, as revealed by computer logs. Fidelity of instruction differentiation was rated by researchers who had observed treatment classrooms biweekly from December to February (control classrooms were rated 0 regardless of classroom observations). Outcomes were assessed with the Woodcock Johnson Tests of Achievement—III, administered in August, January, and May. Because students were nested within classrooms, Hierarchical Linear Modeling was used to analyze fidelity and outcomes. Comparing intervention and control conditions, the intervention resulted in an effect size of 0.25 on student achievement. When both groups were combined, the amount of time teachers used the A2i software accounted for 15% of the variance in student outcomes. Given that all treatment teachers attended professional development but gains were greatest among treatment teachers who most used A2i, the authors were able to conclude that both components were essential to achieving the intervention effect.

Several more examples of methods for linking fidelity with outcomes in the analysis include:

- Conducting an ANOVA both with and without the lowest-fidelity classroom to determine the extent that it impacts overall results[50]
- Determining the correlation of the overall index of fidelity with each student outcome to determine which is most impacted by fidelity[51]
- Determining the correlation of specific implementation behaviors with a single outcome to determine the impact of each[52]
- Conduct a trend analysis of implementation to show how it varies over time, and correlate implementation and outcome trends to show how they covary[25]

Implications for Behavioral Health

Behavioral health interventions range from multisite community programs with multiple components and flexible activities, to one-on-one psychotherapeutic sessions with behaviors

prescribed by the minute.[53] The five-step process for assessing fidelity can be applied to either case or any example between: when an intervention is broad, model-based assessment facilitates the identification of specific and meaningful indicators of fidelity; when the intervention is detailed and specific, the five-step process guides the analysis of fidelity data by combining individual items that represent the same construct. The important distinction made by this process is that *fidelity assessment needs to be model-based, and not simply measures of best practices*. This brings a very specific process for articulating what is to be measured (from constructs to activities to indicators to indices), a process that guides consistent transitions between those pieces—pieces that can be found separately or less systematically in the behavioral health literature but now have been combined in a way that complements the causal model.

Schoenwald and colleagues[9] describe in detail an approach to fidelity assessment in the context of mental health that includes the steps of identifying core components, determining who will measure fidelity and how, collecting fidelity data, and creating a summary fidelity index. The five-step process can be seen as an elaboration on this approach, one which is only enhanced by their in-depth explication of how fidelity can be coded and scored in the clinical setting.

The discussion of intervention fidelity has focused on developing instruments to collect summative data for descriptive and analytic purposes, but fidelity instruments can be used for other purposes that may be particularly relevant for behavioral health interventions. Practical or ethical considerations may lead researchers in health or other fields to employ fidelity measures formatively to provide feedback that can boost implementation.[53] When researchers intervene to manipulate fidelity, the notion of intervention fidelity must be reinterpreted: implementation is no longer the simple effect of assignment to condition, and recorded levels of fidelity and consequent outcomes may not reflect levels and outcomes that would be achieved in the real world without researcher support. Such an experiment might be seen as demonstrating whether the intervention is capable of functioning effectively under ideal conditions, perhaps as a prelude to full-scale deployment (i.e., an efficacy study).

In addition to identifying well-implemented programs, fidelity measures can be used to discriminate between related programs within specific areas of behavioral health intervention. For example, Bond et al.[53] point out that similarities in psychiatric rehabilitation programs have created ambiguity in the literature. They give the example of the DACTS fidelity instrument being used to distinguish between different types of assertive community therapy models[54] and to show when nominally different models may not be significantly different in application.[55]

Other uses of fidelity assessment lie outside of the experimental context entirely. Fidelity measures can be employed by clinicians and program staff in order to track performance and identify opportunities for improvement; by outside evaluators, administrators, or funders to monitor implementation; or to guide the adoption of unfamiliar programs or practices.[9,53] These applications of fidelity assessment largely lie outside of the scope of intervention fidelity as defined here, but many of the same principles apply for developing instrumentation. In fact, instrumentation developed according to the five-step method for use in the context of an experiment can later be adapted and packaged with the intervention as an additional component for tracking and boosting fidelity.

Conclusions

When considering fidelity assessment, researchers must wade through a hodgepodge of approaches and methods, often forced to resort to the sort of unsystematic assessment of the cause that would never be tolerated when assessing outcomes. The five-step process presented herein provides a systematic approach to fidelity assessment that is adaptable to a variety of interventions, experimental designs and

research questions. The researcher applying this process of fidelity assessment begins by answering the question, "Fidelity to what?" and is carried forward from specification to measurement to analysis, with a multitude of tools and options at each step. Admittedly, this approach is best employed from the planning stage of a study,[8] and it is a challenge to try to mold one's results retroactively to the template provided. Nonetheless, each step of the process provides some insight for those with extant data searching for the best methods for analyzing and interpreting it.

In order to achieve the best outcomes for society, researchers must be able to differentiate between interventions that can be effective when implemented as designed and those interventions that are simply ineffective. The methods discussed in this paper assist in that identification within the context of an experiment, both by attributing outcomes to the actual activities that constituted the treatment (as opposed to the theoretical or planned activities) and by allowing for estimates of the relationship between high implementation (where present) and positive outcomes. Yet, identifying quality interventions is not enough: once an intervention is scaled up for real-world implementation, weaknesses in implementation are likely to grow and positive intervention impacts weaken. Thus, in parallel with measuring fidelity, researchers must also attempt to minimize infidelity. A number of papers[3,6,8] have identified factors that either help or hinder implementation, including components of the intervention's logic model and qualities of the intervention context. As this research develops, it should become clear how new interventions can be better designed to improve their chances of being implemented well. Even existing interventions can have their logic models revised to include drivers of implementation, a step that might be taken after demonstrating effectiveness but before scale-up.

Acknowledgments

The authors received support from the Institute of Education Sciences as follows: E.C. Sommer, #R305B04110; M.C. Nelson, #R305B04110; D.S. Cordray, #R305U060002; C.S. Hulleman, #R305B050029 and #144-NL14; C.L. Darrow, #R305B080025. However, the contents do not necessarily represent the positions or policies of the Institute of Education Sciences or the U.S. Department of Education.

Conflict of Interest Each of the authors affirms that there are no actual or perceived conflicts of interest, financial or nonfinancial, that would bias any part of this manuscript.

References

1. Dane AV, Schneider BH. Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review* 1998; 18(1):23–45.
2. McIntyre LL, Gresham FM, DiGennaro FD, et al. Treatment integrity of school-based interventions with children in the *Journal of Applied Behavior Analysis* 1991–2005. *Journal of Applied Behavior Analysis* 2007; 40(4):659–672.
3. Durlak JA, DuPre, EP. Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology* 2008; 41(3):327–350.
4. O'Donnell CL. Defining, Conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research* 2008; 78(1):33–84.
5. Dusenbury L, Brannigan R, Falco M, et al. A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. *Health Education Research* 2003; 18(2):237–256.
6. Fixsen DL, Naoom SF, Blasé KA, et al. *Implementation Research: A Synthesis of the Literature.* FMHI Publication no. 231. Tampa: Louis de la Parte Florida Mental Health Institute, National Implementation Research Network, University of South Florida, 2005.
7. Hulleman CS, Rimm-Kaufman SE, Abry TDS. Construct validity, measurement, and analytical issues for fidelity assessment in education research. In: Halle T, Martinez-Beck I, Metz A (eds.) *Applying Implementation Science to Early Care and Education Programs and Systems: Exploring a New Frontier.* Baltimore, M.D.: Brookes Publishing, in press.
8. Bellg AJ, Borrelli B, Resnick, B, et al. Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the Behavior Change Consortium. *Health Psychology* 2004; 23(5):443–451.
9. Schoenwald SK, Garland AF, Chapman JE, et al. Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research* 2011; 38(1):32–43

10. Carroll C, Patterson M, Wood S, et al. A conceptual framework for implementation fidelity. *Implementation Science* 2007; 2(40):1–9. Retrieved on June 1, 2012, from http://www.implementationscience.com/content/2/1/40.
11. Holland PW. Statistics and causal inference. *Journal of the American Statistical Association* 1986; 81(396):945–960.
12. Institute of Education Sciences. *Education Research Training Grants.* RFA No. IES-NCER-2008–02. Washington, D.C.: US Department of Education, 2007.
13. Cordray DS, Pion GM. Treatment strength and integrity: Models and methods. In: Bootzin RR, McKnight PE (eds). Strengthening Research Methodology: Psychological Measurement and Evaluation. Washington, DC: American Psychological Association, 2006: pp. 103–124.
14. Hulleman CS, Cordray D. Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Intervention Effectiveness* 2009; 2(1):88–110.
15. Borrelli B, Sepinwall D, Bellg AJ, et al. A new tool to assess treatment fidelity and evaluation of treatment fidelity across 10years of health behavior research. *Journal of Consulting and Clinical Psychology* 2005; 73(5):852–860.
16. Lichstein KL, Riedel BW, Grieve R. Fair tests of clinical trials: A treatment implementation model. *Advances in Behavior Research and Therapy* 1994; 16: 1–29.
17. Cordray DS. 2007 Assessing Intervention Fidelity in Randomized Field Experiments. Funded Goal 5 proposal to Institute of Education Sciences.
18. Hulleman CS, Cordray DS, Nelson MC, et al. The state of treatment fidelity assessment in elementary mathematics interventions. Poster presented at the annual meeting conference of the Institute of Education Sciences, Washington, D.C., June 2009.
19. Knowlton LW, Phillips CC. *The Logic Model Guidebook: Better Strategies for Great Results*. Washington, D.C.: Sage, 2009.
20. Chen HT. *Theory-Driven Evaluation.* Thousand Oaks, CA: Sage Publications, 1990.
21. Sidani S, Sechrest L. Putting theory into operation. *American Journal of Evaluation* 1999; 20(2):227–238.
22. Donaldson SI, Lipsey MW. Roles for theory in contemporary evaluation practice: Developing practical knowledge. In: Shaw I, Greene JC, Mark MM (eds). *The Handbook of Evaluation: Policies, Programs, and Practices.* London: Sage, 2006: pp. 56–75.
23. Trochim W, Cook J. Pattern matching in theory-driven evaluation: A field example from psychiatric rehabilitation. In: Chen H, Rossi PH (eds). *Using Theory to Improve Program and Policy Evaluations.* New York: Greenwood Press, 1992, pp. 49–69.
24. Swafford JO, Jones GA, Thornton CA. Increased knowledge in geometry and instructional practice. *Journal for Research in Mathematics Education* 1997; 28(4):467–483.
25. Noell GH, Witt JC, Slider NJ, et al. Treatment implementation following behavioral consultation in schools: A comparison of three follow-up strategies. *School Psychology Review* 2005; 34(1):87–106.
26. Moss M, Fountain AR, Boulay B, et al. *Reading First Implementation Evaluation: Final Report*. Cambridge, MA: Abt Associates, 2008.
27. Shadish WR, Cook TD, Campbell, DT. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* New York, NY: Houghton Mifflin Company, 2002.
28. Cook T. Postpositivist critical multiplism. In: Shotland RL, Marks MM (eds). *Social Science and Social Policy.* Beverly Hills, CA: Sage, 1985, pp. 21–62.
29. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951; 16(3):297–334.
30. Cordray DS. Identifying and Assessing the Cause in RCTs. Instructional session presented at the Institute of Education Sciences RCT Training Institute, Nashville, TN, June 22, 2009.
31. Cronbach LJ, Nageswari R, Gleser, GC. Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology* 1963; 16(2):137–163.
32. Cronbach LJ, Gleser GC, Nanda H, et al. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: John Wiley, 1972.
33. Crocker L, Algina, J. *Introduction to Classical and Modern Test Theory.* New York: Harcourt Brace Jovanovich College Publishers, 1986:527.
34. Brennan LB. Generalizability theory. In: Gierl M (ed). *ITEMS: The Instructional Topics in Educational Measurement Series*. Madison, WI: National Council on Measurement in Education, 1992. Available at: www.ncme.org/pubs/items.cfm Accessed June 18, 2011.
35. Spector PE. *Summated Rating Scale Construction: An Introduction*. Newbury Park, CA: Sage, 1992.
36. Lennon RT. Assumptions underlying the use of content validity. *Educational and Psychological Measurement* 1956; 16(3):294–304.
37. Cronbach LJ. Test validation. In: Thorndike, RL (ed.). Educational Measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971, pp. 443–507.
38. Mosier CI. A critical examination of the concepts of face validity. *Educational & Psychological Measurement* 1947; 7(2):191–205.
39. McGrew JH, Bond GR, Dietzen L, Salyers M. Measuring the fidelity of implementation of a mental health program model. *Journal of Consulting and Clinical Psychology* 1994; 62(4): 670–678.
40. Mowbray CT, Holter MC, Teague GB et al. Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation* 2003; 24:315–340.
41. Abry T, Rimm-Kaufman SE, Hulleman CS. *Using Intervention Core Components to Identify the Active Ingredients of the Responsive Classroom approach.* 2012, manuscript in preparation.
42. Fuchs LS, Fuchs D, Yazdian L, et al. Enhancing first-grade children's mathematical development with peer-assisted learning strategies. *School Psychology Review* 2002; 31(4):569–583.
43. Cordray DS, Pion GM, Dawson M, et al. 2008. The Efficacy of NWEA's MAP Program. Institute of Education Sciences funded proposal.
44. Wilson SJ, Lipsey MW, Derzon JH. The effects of school-based intervention programs on aggressive behavior: A meta-analysis. *Journal of Consulting and Clinical Psychology* 2003; 71:136–149.
45. Tobler NS. Meta-analysis of 143 adolescent drug prevention programs: Quantitative outcome results of program participants compared to a control or comparison group. *Journal of Drug Issues*, 1986; 16:537–567.
46. DuBois DL, Holloway BE, Valentine JC, et al. Effectiveness of mentoring programs for youth: A metaanalytic review. *American Journal of Community Psychology* 2002; 30:157–198.

47. Smith JD, Schneider BH, Smith PK, et al. The effectiveness of whole-school antibullying programs: A synthesis of evaluation research. *School Psychology Review* 2004; 33:547–560.
48. Likert R. A technique for the measurement of attitudes. *Archives of Psychology* 1932; 140:5–53.
49. Connor CM, Morrison FM, Fishman BJ, et al. Algorithm-guided individualized reading instruction. *Science* 2007; 315(5811):464–465.
50. Fuchs LS, Fuchs D, Karns K. Enhancing kindergarteners' mathematical development: Effects of peer-assisted learning strategies. *Elementary School Journal* 2001; 101(5):495–510.
51. Kutash K, Duchnowski A J, Sumi WC, et al. A school, family, and community collaborative program for children who have emotional disturbances. Journal of Emotional and Behavioral Disorders 2002; 10(2):99–107.
52. Ginsburg-Block M, Fantuzzo J. Reciprocal peer tutoring: An analysis of teacher and student interactions as a function of training and experience. *School Psychology Quarterly* 1997; 12(2):1–16.
53. Bond GR, Evans L, Salyers MP, et al. Measurement of fidelity in psychiatric rehabilitation. *Mental Health Services Research*, 2000; 2 (2):75–87.
54. Teague GB, Bond GR, Drake RE. Program fidelity in assertive community treatment: Development and use of a measure. *American Journal of Orthopsychiatry* 1998; 68:216–232.
55. Johnsen M, Samberg L, Calsyn R, et al. Case management models for persons who are homeless and mentally ill: The ACCESS Demonstration Project. *Community Mental Health Journal* 1999; 35:325–346.