Journal of MultiDisciplinary Evaluation
Volume 21, Issue 49, 2025

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

Implementation Fidelity: The Disconnect Between Theory and Practice

Christopher Rhoads
University of Connecticut

Bianca Montrosse-Moorhead
University of Connecticut

Kylie Anglin
University of Connecticut

Catherine Lewis
Northeastern University

Background: This article argues that the measurement of implementation fidelity is impeded by the failure to recognize the existence of competing conceptualizations, rooted in different theoretical traditions, of the concept of fidelity.

Purpose: This paper names competing conceptions of fidelity, highlights the origins of differing conceptions, and uses a case example to illustrate why misalignment between conceptualization and measurement can be problematic in and for practice.

Setting: Not applicable.

Intervention: Not applicable.

Research Design: Not applicable.

Data Collection and Analysis: Not applicable.

Findings: We call competing perspectives context-dependent and context-independent fidelity (i.e., CDF and CIF frameworks, respectively). Different evaluation contexts may be better matched to one or the other of these perspectives. Confusion about how fidelity should be defined in a given funding program or evaluation prevents evaluators from instituting a maximally useful fidelity measurement program. Difficulties inherent to creating high-quality fidelity measures contribute to the problem. We discuss the causes and consequences of this disconnect between fidelity theory and fidelity practice and advance preliminary suggestions for solutions.

Keywords: *implementation fidelity; fidelity frameworks; experimentation; evaluation practice.*

Implementation fidelity plays an increasingly important role in modern evaluations as researchers seek to understand how programs produce positive change (Dane & Schneider, 1998; Hill & Erickson, 2019; O'Donnell, 2008). Yet, evaluators have struggled to adequately conceptualize and measure implementation fidelity since the concept was first introduced (Berman & McLaughlin, 1976; Century et al., 2010; Charters & Jones, 1974). To address these challenges, multiple frameworks have been presented, including those by Century et al. (2010), Dane and Schneider (1998), Cordray and Pion (2006), and Nelson et al. (2012), to name a few. Within these frameworks, multiple definitions of fidelity have been provided. For example, Century et al. (2010) define fidelity as "the extent to which the critical components of an intended program are present when that program is enacted" (p. 202). On the other hand, Hulleman and Cordray (2009) provide the following definition of fidelity: "the specification of a 'gold standard' or basis for comparison—a theory, model, or conception of the educational intervention—to which something is faithful ... [and] how closely the intervention, in practice, met these specifications." (p. 90). Each of these frameworks has helped guide evaluators needing to construct and operationalize fidelity measures. Yet, they also reflect distinct conceptions of fidelity that are, at times, in conflict.

A recent systematic review concluded that, although there are numerous examples of how fidelity is correlated with outcomes, there are still critical under-examined elements of fidelity warranting future work (Hill & Erickson, 2019). In this article, we answer this call by identifying two different and opposing conceptions of fidelity undergirding the rich theoretical base that informs fidelity frameworks. While some frameworks have taken a general approach and identified fidelity dimensions to be measured across all types of program evaluations (e.g., adherence, exposure, quality of delivery, participant responsiveness, and program differentiation; Carroll et al., 2007; Dane & Schneider, 1998), others have taken a more contextualized approach and argued that the fidelity elements to be measured ought to be aligned with local program theories, theories of action, or theories of change (e.g., Cordray & Morphy, 2009; Donaldson, 2007; Funnell & Rogers, 2011). We call these competing conceptualizations context-independent fidelity (CIF) and context-dependent fidelity (CDF), respectively. This article describes the intellectual history of these two competing perspectives, demonstrates how the failure to recognize the distinction between the two perspectives can confuse evaluators and hinder fidelity measurement, and offers a way forward.

We provide an overview of the fidelity literature, highlight the origins of the differing conceptions of fidelity—context-independent and context-dependent—and describe how they map onto different approaches for measuring fidelity. To make visible how the underacknowledged distinction between these two types of fidelity can have practical consequences for evaluations, we describe a case example and discuss how the example illuminates a disconnect between theory and practice when the selected fidelity framework is incompatible with the intervention being evaluated. We conclude the article with suggestions for how evaluators tasked with measuring fidelity can better align the practice of fidelity measurement with theory.

A Brief Review of Fidelity Frameworks

To measure fidelity, one needs to first articulate what is meant by the term "fidelity." Other authors have noted that the way in which "implementation fidelity [is] conceptualized and measured continues to vary ... making interpretation ... nebulous" (Meyers & Brandt, 2014, p. 14). This has important implications, as it goes to a central and key question—what do we mean by "fidelity"? To provide insight, we elaborate the intellectual history of what we call context-dependent and context-independent fidelity. In what follows, we describe two dominant and competing conceptualizations of fidelity—a distinction that has not previously been acknowledged in published literature.

Context-Independent Approach to Fidelity

Dane and Schneider introduced a fidelity framework in 1998. Inspired by and drawing on a critical literature review, which examined 162 outcome evaluations (of maladaptive behavior, social, or academic interventions) published between 1980 and 1994, Dane and Schneider argued that fidelity should be defined as "the degree to which programs were implemented as planned" (Dane & Schneider, 1998, p. 23). They further argued that any fidelity study should include an empirical examination of five dimensions: adherence, exposure (conjointly called dosage), quality of delivery, participant responsiveness, and program differentiation. *Adherence* is how closely the delivery of the program followed what was outlined by the program developers. *Exposure* refers to such elements as the frequency of implemented sessions, the length of said sessions, and the number of sessions. *Quality of delivery* focuses on the implementer in terms of his or her attitude, preparedness, and enthusiasm. *Participant responsiveness* is similar to quality of delivery in terms of the focus on attitude and enthusiasm, yet it focuses on the participant response to the intervention rather than the implementer. The final component, *program differentiation*, is "a manipulation check" which examines the difference between business-as-usual and the treatment program to ensure there was no "unintentional spread" of treatments (Dane & Schneider, 1998, p. 45). Table 1 summarizes the key elements of this framework.

Dane and Schneider's (1998) framework, we contend, is rooted in the idea that the conceptualization of fidelity should be context-independent (i.e., it is a CIF framework). We use this language because Dane and Schneider argue that these same five dimensions can and should be used to assess fidelity across any type of intervention. This argument can be traced back to literature from scientific management and behaviorist perspectives in psychology and medicine (Berman & McLaughlin, 1976; Havelock, 1969; House et al., 1972; Rogers, 1995). Extended to practice, it means that all evaluations of educational interventions, public health interventions, psychosocial interventions, and so on would use these same five components to capture fidelity, even if the context and nature of the interventions are vastly different.

Dane and Schneider's (1998) framework has been popular and influential (Hill & Erickson, 2019). To date, it has been cited over 2,300 times in published work (Google Scholar, December 2024). We suspect its popularity is due to the clarity of the framework and the straightforward guidance it supplies for operationalizing fidelity.

Other evaluators have built upon Dane and Schneider's original framework (e.g., Century et al., 2010; Hill & Erickson, 2019; Mowbray et al., 2003). One recent example is the framework provided by Century et al. (2010). Their motivation, like Dane and Schneider's, was to develop an approach to fidelity measurement that can work for multiple programs. Thus, their framework reorganizes Dane and Schneider's five dimensions, adds a sixth one (i.e., educative), and explicates each dimension by drawing on other work that illuminates the importance of program structure and important human interactions during program delivery (e.g., Mowbray et al., 2003), which they refer to as the "structure" and "process" dimensions of fidelity. Aligned with their motivation to create a fidelity framework that would work across multiple programs, they include a list of fidelity indicators aligned to each dimension of their framework, most of which they label as common across all types of mathematics and science programs. By creating a comprehensive list of fidelity indicators, Century et al. (2010) helped move forward context-independent fidelity measurement.

Table 1. Components of the Context-Independent and Context-Dependent Fidelity Frameworks

Framework orientation (authors)	Fidelity definition	Dimensions & sub-dimensions	Dimension definition
Context-Independent (Dane & Schneider, 1998)	"The degree to which specified procedures are implemented as planned" (p. 23).	Adherence	"The extent to which specified program components were delivered as prescribed in program manuals" (p. 45).
		Exposure	"An index that may include any of the following: (a) the number of sessions implemented; (b) the length of each session; or (c) the frequency with which program techniques were implemented" (p. 45).
		Quality of delivery	"A measure of qualitative aspects of program delivery that are not directly related to the implementation of prescribed content, such as implementor enthusiasm, leader preparedness, global estimates of session effectiveness, and leader attitudes toward program" (p. 45).
		Participant responsiveness	"A measure of participant response to program sessions, which may include indicators such as levels of participation and enthusiasm" (p. 45).
		Program differentiation	"A manipulation check that is performed to safeguard against the diffusion of treatments, that is, to ensure that the subjects in each experimental condition received only planned interventions" (p. 45).
Context-Dependent (Hulleman & Cordray, 2009; Nelson, Cordray, Hulleman, Darrow, & Sommer, 2012)	"The extent to which an intervention's core components have been delivered as prescribed and differentiated from the comparison condition" (Nelson et al., 2012, p. 375).	Change model (i.e., program theory)	Represents "an understanding of the theoretical basis for the intervention's form and function.... 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" (Nelson et al., 2012, p. 381).
		Core components	Represents "elements that are unique to the intervention (as compared to a counterfactual condition) and are essential to achieving its effects" (Nelson et al., 2012, p. 381). "Components are simply the major constructs represented in the change model; components are often multidimensional" (Nelson et al., 2012, p. 384).
		Subcomponents	"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" (Nelson et al., 2012, p. 385).
		Facets	"The specific behaviors, events or resources that constitute the implementation of a subcomponent" and "not to be confused with the facets of generalizability theory" (Nelson et al., 2012, p. 385). One or more facets may be appropriate in a given study.
		Indicators	Provides "evidence of the degree to which each facet is implemented" (Nelson et al., 2012, p. 385). Represents what is to be measured in the evaluation. "Any indicator that

Framework orientation (authors)	Fidelity definition	Dimensions & sub-dimensions	Dimension definition
Context-Dependent (Hulleman & Cordray, 2009; Nelson, Cordray, Hulleman, Darrow, & Sommer, 2012			cannot be tied back to a core 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.” (Nelson et al., 2012, p. 385). One or more indicators may be used for each facet.
		Indices	Also helps provide “evidence of the degree to which each facet is implemented” (Nelson et al., 2012, p. 385). Represents how indicators “are to be directly measured” (Nelson et al., 2012, p. 385).
		Absolute fidelity index	Measure of “whether there was an absolute, or maximum, level of fidelity from which to compare participants’ responses” (Hulleman & Cordray, 2009, p. 96).
		Average fidelity index	Measure of the “mean levels of response quality in each condition as an indicator of treatment receipt” (Hulleman & Cordray, 2009, p. 96).
		Binary complier fidelity index	Measure of “whether participants had received the treatment (or not)” (Hulleman & Cordray, 2009, p. 97).
		Achieved relative strength	Measure of “the degree of discrepancy between what should have been implemented and what was actually implemented” (Hulleman & Cordray, 2009, p. 90).

Note. While we also discuss a different framework (Century et al., 2010) within the text, in this table we only describe features of the original frameworks that we believe are *most* aligned with the CIF and CDF conceptualizations of fidelity

Context-Dependent Approach to Fidelity

In the late 1970s Rossi and Sechrest separately became concerned about the neglect of measurement of the "treatment" in evaluation questions and resulting studies (Chen & Rossi, 1980; Rossi et al., 2004; Sechrest & Redner, 1979). While questions of impact and outcomes were important, they argued, so too was the concept of treatment strength and treatment integrity. Sechrest and Redner (1979) defined *treatment strength* as the a priori hypothesis about how strong a treatment needed to be, compared to business-as-usual, to produce desired effects, and *treatment integrity* as the degree of alignment between the planned versus enacted treatment. Later, Lipsey et al. (1985) showed that over 70% of studies they reviewed offered no information on treatment strength or integrity or, if they did, only offered a general, nonempirical statement along the lines of "Neither treatment strength nor treatment integrity was measured." Evaluators began to push forward ideas about how to address this problem, drawing inspiration from the concerns-based adoption model (Hall & Loucks, 1977) and socioecological theory (Bronfenbrenner, 1979). One proposal put forth to define both treatment strength and treatment integrity was to use program theories (Bickman, 1987; Rossi et al., 2004).[1] For example, a program theory could be used to define ideas about how strong a treatment needs to be, through the use of benchmarking, and to specify what mechanisms differentiate an intervention from business-as-usual. This line of thinking eventually led to an expectation that program evaluators and developers should articulate program theories that embed both a theory of change and a theory of action (Funnell & Rogers, 2011).

Building upon this early work, Cordray argued that treatment has to be "ruled in" as an explanation for causes, further lending support to the need for evaluations that used program theory. This was an impetus for Cordray's fidelity framework to improve the measurement of both treatment strength and treatment integrity (e.g., Cordray, 1989; Cordray & Pion, 2006; Orwin et al., 1998). Recent work with colleagues has focused on the creation of indices of implementation (Cordray et al., 2013; Hulleman & Cordray, 2009; Nelson et al., 2012).

Because the theory of change is unique to the specific intervention for which fidelity is being measured, the program theory–based model framework can be thought of as a context-dependent approach to fidelity (CDF framework). In the practical sense, this means evaluators evaluating interventions first need to develop a program theory that is unique to what is being evaluated and then develop fidelity measures aligned with "elements that are unique to the intervention (as compared to a counterfactual condition) and are essential to achieving its effects" (Nelson et al., 2012, p. 381). Thus, both context and the nature of the interventions play an important role in fidelity measurement.

Cordray's work has also been well received (Abry et al., 2015). Across all of Cordray's fidelity publications, the framework has been cited over 1,050 times, suggesting that others are using this approach in practice (Google Scholar, December 2024). However, as we note elsewhere, the CDF framework appears to be utilized much less often than the CIF framework in practice (Montrosse-Moorhead et al., 2016). We suspect that one reason the CDF framework has proven to be less popular, as compared to Dane and Schneider's framework, is that it requires more time and resources to develop a program theory unique to each intervention and subsequently define fidelity measures aligned with the core components of the program theory. We elaborate on the measurement challenges posed by the CDF framework later in this article and we provide a summary overview of key distinctions between the CIF and CDF frameworks in Table 1.

Case Example: Japanese Lesson Study

Having briefly reviewed the literature on fidelity, we now describe an evaluation of a particular intervention—Japanese lesson study—as a real-life example of how using a fidelity framework incompatible with the intervention being evaluated has implications for practice.[2] The case presented here is a simulation-of-practice case, meaning its purpose is to help readers engage in meta-evaluative learning by critically examining and analyzing the case for the purpose of illuminating key ideas (Ensminger et al., 2021). In the Japanese lesson study case, the key idea is that there is more than one framework to measure fidelity, and it is

[1] Within the evaluation literature various terms are used to refer to a model of how a program is, in theory, intended to work; for example, program theory, theory of change, intervention theory, and so on (Funnell & Rogers, 2011).

[2] Lesson study is a type of collaborative professional development used by teachers.

important for evaluators to choose a fidelity framework using this knowledge.

The case example is from an evaluation of a professional development intervention that aimed to have teams of teachers use Japanese lesson study to build their knowledge and teaching related to fractions. Lesson study is a form of professional learning that has been in use for more than a century in Japan (NIER, 2011), and it has spread to many other countries (WALS, 2012) since the first English-language accounts appeared (Lewis & Tsuchida, 1997; Stigler and Hiebert, 1999). Lesson study has spread rapidly and widely in the United States since its introduction in the late 1990s (Akiba et al., 2014; Hill, 2011).

Japanese lesson study includes four interrelated phases. When implementing Japanese lesson study, a small team of educators conducts a four-part cycle of inquiry: (1) Study particular content; (2) plan a "research lesson" and unit that reflects the team's thinking about the optimal teaching of that content; (3) one team member teaches the research lesson to students, while others observe students and collect data on student responses; and (4) share, discuss and reflect on the data collected and draw implications for future teaching of the topic and of the content more broadly (Lewis & Hurd, 2011; Wang-Iverson & Yoshida, 2005). Research lessons are so named because lesson study team members (and sometimes additional colleagues) observe, document, and discuss them, focusing on student responses. In this case example, each teacher team included at least one grade 3 or 4 teacher designated for study.

Protocols support some parts of the lesson study process. For example, the observation protocol asks observers not to help or interact with students, and the post-lesson discussion protocol specifies that the teacher of the lesson has the chance to speak first and that observers provide data-based comments rather than inferences. However, educators are expected to structure many parts of the lesson study process themselves—for example, to choose their team, the lessons to study, how to teach them to their students, how to plan activities around these lessons, etc. They are also free to choose how often the group should meet, the timing of those meetings, and the way meetings should be organized. Indeed, the philosophy of lesson study, or its theory of change, is that teachers need to be given control over how to structure their lesson study groups in order for the professional development to be meaningful and to produce the intended effects. Moreover, while impacts on students are an important part of Japanese lesson study's theory of change, the specifics of the anticipated impacts are context dependent. In this case example, students were expected to increase their understanding of fractions because of their teacher's participation in fractions lesson study.

The lack of a defined protocol for how teachers should structure their lesson study groups led the research team to believe that it would not be very useful to measure fidelity. However, the evaluation was funded by the Institute of Education Sciences (IES) in response to a request for proposals (RFP) issued by the agency. The RFP required that grantees include "measures of implementation fidelity" as part of the evaluation but provided little additional guidance regarding the nature of those measures (Institute of Education Sciences, 2013). Because the research team was familiar only with Dane and Schneider's (1998) fidelity framework, while they recognized that measuring fidelity was a requirement of funding, they were unsure of how to measure fidelity in a meaningful way. Specifically, they wondered, without a codified protocol, to what exactly were teachers supposed to be demonstrating fidelity? It seemed to the evaluation team that it might not even be appropriate to measure fidelity at all. Adherence did not make sense, as each team was free to structure many parts of the lesson study process. Exposure (dosage) could possibly be measured, but focusing on the number of meetings, meeting length, and meeting frequency did not really capture what was unique or important about Japanese lesson study. Furthermore, it was expected that control group teachers would also be meeting to do lesson planning, as this is common practice in schools. So, exposure was not unique to the treatment condition. A case could be made for measuring quality of delivery, but there was not an expectation that teacher enthusiasm and attitudes toward teaching fractions or teacher preparedness would be different between the treatment and control groups. The same was true for participant (student) responsiveness. There was no expectation that student enthusiasm would be different. Although it was possible, the intervention did not require that student participation would be different in the treatment and control groups.

Further, even if sensible measures could be identified, what purpose, if any, would the resulting fidelity data serve in the evaluation? While a case could be made for some parts of Dane and Schneider's (1998) fidelity framework as described above, the research team was not convinced that criteria such as adherence, exposure, quality of delivery, and participant responsiveness would focus on the aspects of Japanese lesson study that made it work and made it distinct from other forms of teacher professional development. Nonetheless, because fidelity measurement was required by IES and because Dane and Schneider's (1998) fidelity framework was the only one of which they were aware, the research team collected and reported on adherence, exposure, quality of delivery, and participant responsiveness.

The Tensions That This Case Illuminates: Coherence Between Fidelity Approach, Evaluation Objectives, and Program Philosophy

Because the existing literature has not recognized the distinction between the two visions of fidelity—context-dependent and context-independent—principal investigators (PIs) can be confused when responding to calls from funding agencies to incorporate fidelity into an evaluation. PIs may not be familiar with both conceptualizations of fidelity and so may approach fidelity measurement from a perspective that does not fit the evaluation context. In fact, this is exactly what happened in the Japanese lesson study case example. Difficulties arose for the research team in this case because of a failure to recognize the two competing and incommensurable definitions of fidelity. The lesson study program philosophy required a CDF framing. However, the research team's previous experience led them to associate the term "intervention fidelity" with the CIF frame. They assumed that the funding agency expected CIF measurement using Dane and Schneider's (1998) fidelity framework.

We claim that the Japanese lesson study case example is not unique. The failure to recognize the distinction between context-dependent and context-independent fidelity has impoverished the practice of fidelity measurement. Specifically, this failure has prevented the context-dependent framework from being used to its full potential in evaluations where its use is needed. Instead, there is a wide gulf between the rich theoretical conceptualizations of fidelity that exist in the evaluation literature and the practice of fidelity measurement that occurs in most evaluations. In other work, we present the results of a systematic review of fidelity studies that support this claim (Montrosse-Moorhead et al., 2016). This review finds that only 13.3% of studies measuring fidelity adopt a CDF frame. We conclude that the CDF framework is under-utilized in applied evaluations.

Of particular concern, funding agencies that promote the measurement of fidelity through their RFPs have not historically recognized these competing definitions and so have not provided adequate guidance to grantees, nor adequate time and funding, to allow robust measurement of fidelity. This is not a moot point, as policy changes have intensified calls to include implementation studies in federally funded evaluations across a variety of fields (cf. Alcohol, Drug Abuse and Mental Health Services Administration Reorganization Act of 1992; Education Sciences Reform Act of 2002).

Interest in the concept of fidelity shows no signs of abating anytime soon. If we aim to understand and provide evidence of what works, for whom, and under what conditions, then greater attention to better aligning fidelity theory and practice is needed. In what follows, we advance preliminary suggestions for how to achieve this alignment.

Suggestion 1: Articulate Fidelity *to What?*

To measure fidelity, one needs to articulate fidelity *to what*. In the Japanese lesson study example, the evaluation team imagined that the "*what"* was a set of standardized and universal protocols and procedures that teachers should follow that would add up to faithful implementation of lesson study. However, implementation of the protocols for planning, observing, and discussion misses the essence of the lesson study process, in which teachers are building a team structure that allows them to share and build knowledge about instruction. Teachers can implement the surface features of lesson study (protocols) without building the underlying changes that enable it to work (changes in knowledge, beliefs about instruction, habits of noticing student thinking, etc.). This is a problem that often arises when fidelity is equated solely with adherence.

The lesson study evaluation team is not alone in equating fidelity with the adherence dimension of the CIF framework. Our systematic review (Montrosse-Moorhead et al., 2016) shows that *adherence* is by far the most frequent type of fidelity measured in evaluations. Approximately 55% of the studies reviewed used adherence as their sole fidelity measure. Yet, standard indicators of adherence (such as the amount of time a teacher spent implementing a program) are unlikely to provide the kind of rich information that developers need to further refine and improve interventions, especially interventions that are nascent and not prescriptive (see Suggestion 3 below).

Regardless of whether a CIF or CDF framework is used, we concur with others that good fidelity measures must be able to distinguish positive from negative infidelity (Century et al., 2010; Munter, et. al., 2014). Most existing adherence measures are not well suited to make this distinction, because they are grounded in a concept that does not allow for any changes to be made. Any deviation from procedure is deemed infidelity, even if the deviation might be consistent with the program theory of change. The evaluation and measurement community have long argued that consequential validity and attending to unintended/unanticipated outcomes/consequences, which can be positive, is important (Bamberger et al., 2016; Jabeen, 2016; Messick, 1995; Scriven, 1973; Shepard, 1997). Acknowledging the possibility of positive infidelity (Munter et al., 2014) or acceptable adaptation (Century et al., 2010) is the extension of this argument to the area of fidelity.

Focusing attention on the "Fidelity to what?" question is a useful way of ensuring that principal investigators and study staff are approaching the task of fidelity measurement using an appropriate frame, which in many cases means moving beyond a purely adherence-based view of fidelity. Specifically, it requires study staff to think through and take a position on whether positive infidelity and/or acceptable adaptations are, in theory, compatible with the operationalization of the selected fidelity measurement frame. Thus, thinking through the "Fidelity to what?" question is an important, often overlooked, part of planning for fidelity measurement in evaluation studies.

Suggestion 2: Key Stakeholders Can Help Inform Fidelity Framework Selection

Returning to our example, while the set protocols associated with lesson study capture only a small part of the process, Japanese lesson study experts agree that there are a set of core values and associated practices that characterize a well-conceptualized and well-functioning lesson study group. For instance, one member of the advisory board for the evaluation was formerly a schoolteacher in Japan for many years and was intimately familiar with lesson study as practiced in that country. During conversations about fidelity, he was able to clearly articulate activities and interactions that should be occurring in a lesson study group that was demonstrating fidelity to the underlying theory. He argued that there is a particular manner in which the teachers should go about researching the topic, go about teaching or observing the research lesson, and go about extracting information and supporting teacher development post–research lesson. His contributions to the discussion demonstrated that experts may be able to articulate both a theory of change and measurable core components linked to that theory of change, even for interventions that lack rigid protocols. Additionally, his expertise demonstrated that even for a nonprescriptive intervention like Japanese lesson study it is possible to distinguish well-functioning lesson study groups both from business-as-usual practice and from lesson study groups that are not appropriately following the lesson study model.

There was also agreement among many members of the advisory board that many of the activities that get called Japanese lesson study in current practice in the United States are *not* consistent with the core principles of lesson study as practiced in Japan. In other words, while lesson study does not specify a set of protocols for teachers to follow, lesson study does entail a set of expectations about the sorts of activities teachers will engage in and the sorts of interactions that will occur in the context of a lesson study group. Furthermore, these activities and interactions are both measurable and distinct from what would be expected in standard practice. In other words, it should be possible to obtain good quantitative information about the extent to which the lesson study intervention encourages certain theoretically desirable behaviors. Measuring these behaviors is the essence of fidelity measurement consistent with the CDF framework.

In short, advisory board meetings for the lesson study project revealed rich information about how experts conceptualize lesson study that could have been used to inform fidelity measurement for the project. Unfortunately, by the time the advisory board meetings took place the basic structure for fidelity measurement was locked in by the funding mechanism and unchangeable. However, the next sections explain how future evaluations can establish a shared understanding of fidelity in order to implement a fidelity protocol consistent with the CDF framework.

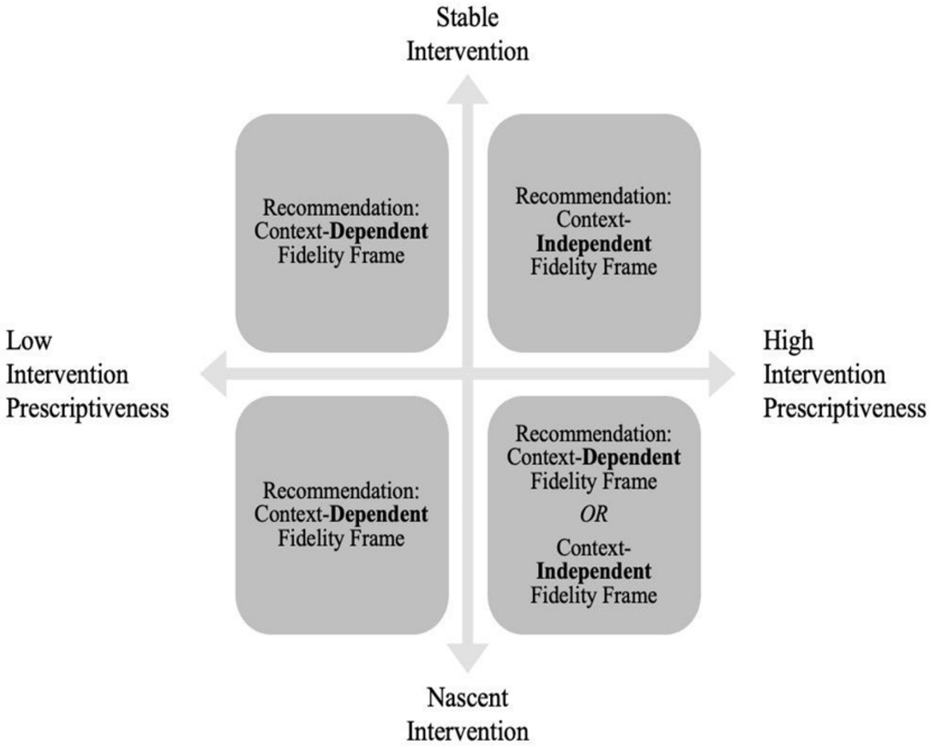
Suggestion 3: Understanding Intervention Prescriptiveness and Stage of Development are Necessary but not Sufficient for Choosing a Fidelity Framework

Program developers, funders, researchers, and evaluators must be more transparent about whether a CDF or a CIF frame is being used (or should be used) for a particular evaluation. Particularly in the context of impact evaluations where evaluation objectives are causal, we suggest that when making a fidelity measurement plan evaluators should consider the prescriptiveness of the intervention and the stage of development.

Figure 1 presents a heuristic for making decisions about which fidelity framework to use. When the intervention is not prescriptive (meaning the intervention theory of change leaves many key decisions up to implementers), and regardless of whether the intervention is nascent or stable, we argue that a CDF framework is needed to identify the core components that make the intervention work and differentiate it from the control group.

However, for interventions that are highly prescriptive (see, for example, Fuchs et al., 1990), the stability of the intervention differentiates when to use which fidelity framework. For established and prescriptive interventions, we argue that a CIF framework is needed precisely because the intervention is so structured, and implementation should look the same regardless of context. For nascent prescriptive interventions, either frame can be used. A CIF framework may be preferred to ensure that deviations from the prescribed intervention are appropriately captured.. Alternatively, a CDF framework could be used to help inform program development and planning. Adopting a CDF framework at this stage has the added benefit of allowing implementers the flexibility to adapt the program theory to different circumstances without characterizing these adaptations as “infidelity.” That is, it allows evaluators to account for positive infidelity (Munter et. al., 2014) or acceptable adaptations (Century et al., 2010).

Figure 1. Heuristic for Making Decisions About Which Fidelity Framework to Use in Causal-Oriented Evaluations Based on Intervention Prescriptiveness and Intervention Stage of Development


Stable Intervention
Recommendation: Context-**Dependent** Fidelity Frame
Recommendation: Context-**Independent** Fidelity Frame
Low Intervention Prescriptiveness
High Intervention Prescriptiveness
Recommendation: Context-**Dependent** Fidelity Frame
Recommendation: Context-**Dependent** Fidelity Frame *OR* Context-**Independent** Fidelity Frame
Nascent Intervention

While thinking along these two dimensions of fidelity should help improve fidelity measurement practice, this alone will not be sufficient to allow fidelity measurement to reach its full potential in evaluation. One clear impediment to consistent, high-quality fidelity measurement is the need to invest substantial time and money in measure development. To measure distal outcomes, it will often be possible to utilize pre-existing measures. For instance, in educational evaluations the distal outcome of interest is often academic achievement. Achievement is easily measured with pre-existing standardized tests. Even better, participating schools often collect this data anyway, substantially reducing the data collection burden for both evaluators and students. In studies of behavioral health, there are likely to be pre-existing instruments that can measure the relevant health-related behaviors. An evaluation clearly saves both time and costs by using pre-existing measures. Another benefit to pre-existing measures is that they will often have been subjected to prior psychometric work to explore validity and reliability.

It is much more difficult for evaluators to use pre-existing measures to engage in fidelity measurement work. A good fidelity measure of intervention core components is one that can capture the unique features that distinguish a particular intervention from other practices. These measures need to be specific to the particular intervention under study. In fact, our review (Montrosse-Moorhead et al., 2016) found that roughly half of fidelity studies (49.7%) reviewed required the development of new fidelity measurement tools, and another 19% needed to adapt existing instruments. Empirically, it is unlikely that pre-existing measures can be used for fidelity measurement without significant adaptation. It is even more unlikely that previous psychometric work will be applicable. Our review found that under 5% of studies provided validity evidence for the fidelity measures that were used, and only 1% of studies measuring core components presented such evidence.

We hypothesize that one reason the CIF framework has become prominent is that the dimensions of fidelity that it articulates are easier to measure than the dimensions suggested by the CDF framework. This is particularly true of exposure and adherence. In educational evaluations, for example, exposure can be measured by the amount of time that a teacher spends per week on a certain type of instruction. Or a social work program may measure adherence by the number of times per month or year that clients receive visits from a social worker. Adherence can be similarly straightforward to measure; for example, as described by Atul Gawande in *The Checklist Manifesto*, an intervention may dictate that a surgeon follows a prescribed series of steps in preparing for surgery (Gawande, 2014). This type of adherence can easily be measured with simple and inexpensive modalities such as checklists or logs.

On the other hand, the key ingredients that theoretically make a given (nonprescriptive) intervention effective usually have little to do with such measures. While lack of adherence (e.g., lack of delivery of lessons in a classroom) is likely to impede the goals of an intervention and will be captured by these measures, there is much more happening that is more nuanced and not captured. We suspect that less attention is given to the CDF framework not because evaluators find it less compelling, but because it is harder to measure.

If the recent reproducibility crisis in medicine and psychology has taught us anything, it is that all knowledge must be viewed as contingent. Today's empirical findings are only temporary stops on the path to truth, liable to be overturned by future results. Deviations from protocols should not universally be treated as negative. We believe that moving fidelity practice away from an adherence-based vision of fidelity will help to ensure this goal is reached. The use of fidelity measures specifically tailored to the theory of change underlying the intervention being evaluated (if using a CDF frame) or fidelity measures specifically tailored to the universal fidelity indicators (if using a CIF frame) is a key step (Century et al., 2010; Munter et al., 2014).

Suggestion 4: Recent Developments in Data Science May be Helpful for Reducing Costs Associated with Context-Dependent Fidelity Measurement

The above arguments lead to a conclusion that high-quality fidelity measurement will require a substantial investment of time and resources by evaluation teams. The Japanese lesson study project was fortunate to have a member of the advisory board who could clearly articulate the basic philosophy of the approach, as well as what a particular instantiation of that approach should look like. This advantage will not always be present. In most cases, just to develop appropriate instruments, evaluators will need to spend substantial time and resources interviewing key stakeholders, developing theories of change, and identifying the unique core components (be they practices or principles) that will inform fidelity measure development. Program developers will need to be actively engaged to assist with the development of a program's theory of change. Key implementers will need to provide information about possible impediments to implementation and possible positive local adaptations for which evaluators should be on the lookout.

Even once instruments have been developed, fidelity measurement remains resource intensive. In certain circumscribed instances, surveys, checklists or narrative time logs may be appropriate. However, often direct observation (either in person or using video) will be necessary. In this case an appropriate and usable observation protocol needs to be developed. Raters need to be trained to recognize desirable and undesirable behaviors. Preliminary work observing business-as-usual practice may be needed to accurately characterize the ways in which intervention core components differ from standard practice. Additional work will need to be done to ensure that the privacy of participants is protected. Evaluators should engage in cognitive interviews with program developers and expert implementers to extract core components and recognize positive adaptations.

Recent work in natural language processing (NLP) may provide a useful way forward in some situations (Anglin et al., 2021). Given NLP techniques are designed to aid humans in analyzing large amounts of language data, they are particularly useful in the evaluation of language-based treatments—treatments where the core components focus on spoken or written interactions. Japanese lesson study qualifies as a language-based intervention (key components focus on the conversation between teachers during the lesson debrief). Other language-based interventions include therapy (which involves verbal interactions between therapists and clients), curricula (which are often delivered verbally by teachers), and many professional development interventions (which involve interactions between professionals).

One recently established approach to measuring fidelity for highly prescriptive treatments (i.e., within a CIF framework) is measuring the semantic similarity between the language used by implementers (e.g., a teacher implementing a curriculum) and a (possibly scripted) treatment protocol (Anglin et al., 2021). For example, consider a curriculum which asks teachers to provide students with specified definitions of new vocabulary words. Semantic similarity could be used to measure the similarity of the definitions provided by the teachers to the definitions specified in the curriculum's materials. Importantly, semantic similarity methods can be modified to be robust to arbitrary differences in language that do not change the meaning of the interaction (Gomaa, 2013). Nonetheless, this is an approach to measuring fidelity which leaves little room for positive infidelity or acceptable adaptation; a high semantic similarity score is assumed to be better than a lower semantic similarity score. However, it is worth noting that the approach can also be used to explore variations in implementation while maintaining an agnostic attitude towards the degree of similarity (Anglin et al., 2021).

Because of the lower cost of many CIF-associated measures (such as adherence) the greatest gains from using NLP to measure fidelity likely come within a CDF framework. The traditional approach to measuring fidelity within a CDF framework requires substantial effort even after identifying the key components; commonly, researchers need to hire trained observers to identify instances of the components in each treatment session. Text classification offers a more scalable solution; given transcripts of a random sample of treatment sessions, the researcher can train a machine learning model to replicate the work of human observers. For example, in one of the first applications of NLP for the measurement of fidelity, psychologists trained a text classifier to detect therapist reflections, a core "active ingredient" in motivational interviewing treatments (Can et al., 2016). A similar approach could have been used in the Japanese lesson study experiment. Research suggests that one key ingredient of the lesson study is the discussion of evidence of student learning during the debrief and discussion (Perry & Lewis, 2009). A classifier could have been trained to identify instances of this activity within transcriptions of the lesson debrief.

A key advantage of using NLP to measure fidelity is its scalability; while the classifier may require hundreds, or in some cases thousands, of examples to learn from, after training and validation, it may be used again and again at limited cost. Further, if the classifier were trained on examples spanning several lesson contexts (e.g., lessons about levers, pendulums, subtraction, etc.), the classifier may be generalizable enough to be used in new experiments (Jensen et al., 2020), so long as the program theory of change remains the same.

Still, classifiers are not always a low-cost method of measuring fidelity, at least at the start. They often require many hand-labeled transcripts to learn from, as well as ongoing validation and monitoring to ensure that the treatment's core components are reliably identified in the setting in which the classifier is applied. Later, however, start-up costs may be reduced as researchers can build on pretrained large language models, adapting high-performing models to new tasks by providing the classifier with just a few examples and non-examples, eliminating the need for a large number of labeled transcripts for training (Brown et al., 2020). Of course, the output of such models will still require validation, comparing model output to expert ratings of fidelity. To our knowledge, large language models have not yet been used for fidelity measurement, but we hope that these recent developments encourage evaluators to experiment with new approaches within a CDF framework.

Conclusion

While funders often require that fidelity be measured, there seems to be little insistence that fidelity be measured well. Funders who desire evaluations to produce useful information about intervention fidelity will need to encourage grantees to invest substantial time and money into the development of intervention fidelity measures. Granting agencies will also need to provide adequate resources for this effort. This does not seem to be happening at present. A dedicated funding stream for the development of fidelity measures may be required.

However, funding alone will not be sufficient. There also needs to be recognition by those conducting evaluations measuring fidelity of the need for high-quality measurement expertise on the evaluation team. An examination of LaVelle's (2018) *Directory of Evaluator Education Programs in the United States* illustrates that such expertise is unlikely to exist without specific recruitment efforts. Less than half of evaluation programs require a measurement course, less than a quarter require a survey design course, and only one offers a course devoted exclusively to implementation evaluation.

While we are pessimistic about the current state of practice in fidelity measurement, we are optimistic about the prospects for a better tomorrow. With recent advances in technology evaluators have the tools and the data available to create high-quality measures of implementation fidelity. IES has taken steps to clarify how it wants its grantees to measure fidelity, both in its RFPs (IES, 2023) and in the "implementation" and "core components" elements of its standards for excellence in educational research (SEER) principles (IES, 2021). If other agencies follow suit and also begin to provide necessary resources, robust fidelity measurement matched to the evaluation context may one day be a standard feature of evaluations.

References

Abry, T., Hulleman, C. S., & Rimm-Kaufman, S. E. (2015). Using indices of fidelity to intervention core components to identify program active ingredients. *American Journal of Evaluation*, 36(3), 320–338. https://doi.org/10.1177/1098214014557009

Akiba, M., Ramp, L., & Wilkinson, B. (2014). *Lesson study policy and practice in Florida: Findings from a statewide district survey*. Florida State University.

Alcohol, Drug Abuse and Mental Health Services Administration Reorganization Act of 1992, *Public Law 102-321* (p. 141).

Anglin, K. L., Wong, V. C., & Boguslav, A. (2021). A natural language processing approach to measuring treatment adherence and consistency using semantic similarity. *AERA Open*, *7*. https://doi.org/10.1177/23328584211028615

Bamberger, M., Tarsilla, M., & Hesse-Biber, S. (2016). Why so many "rigorous" evaluations fail to identify unintended consequences of development programs: How mixed methods can contribute. *Evaluation and Program Planning*, *55*, 155–162. https://doi.org/10.1016/j.evalprogplan.2016.01.001

Berman, P., & McLaughlin, M. W. (1976) Implementation of educational innovation. *The Educational Forum*, *40*(3), 345–370.

Bickman, L. (1987). The functions of program theory. *New Directions for Program Evaluation*, *33*, 5–18. https://doi.org/10.1002/ev.1443

Bronfenbrenner, U. (1979). Contexts of child rearing: Problems and prospects. *American psychologist*, *34*(10), 844.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Can, D., Marín, R. A., Georgiou, P. G., Imel, Z. E., Atkins, D. C., & Narayanan, S. S. (2016). "It sounds like...": A natural language processing approach to detecting counselor reflections in motivational interviewing. *Journal of Counseling Psychology*, *63*(3), 343. https://doi.org/10.1037/cou0000111

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation*

Science, *2*(40) https://doi.org/10.1186/1748-5908-2-40

Charters, W. W., & Jones, J. E. (1974, February). *On neglect of the independent variable in program evaluation*. University of Oregon, Project MITT.

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation, 31(2)*, 199-218.

Chen, H.-T., & Rossi, P. H. (1980). The multi-goal, theory-driven approach to evaluation: A model linking basic and applied social science. *Social Forces*, *59*, 106–122.

Cordray, D. S. (1989). Optimizing validity in program research: An elaboration of Chen and Rossi's theory-driven approach. *Evaluation and program planning*, *12*(4), 379–385.

Cordray, D. S., & Morphy, P. (2009). Research synthesis and public policy. *The handbook of research synthesis and meta-analysis*, *2*, 473–494.

Cordray, D. S., & Pion, G. M. (2006). Treatment strength and integrity: Models and methods. In R. Bootzin & P. McKnight (Eds.), *Contributions of Lee Sechrest to methodology and evaluation*. APA.

Cordray, D. S., Pion, G. M., Brandt, C., & Molefe, A. (2013). *The impact of the measures of academic progress (MAP) program on student reading achievement*. Society for Research on Educational Effectiveness.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, *18*, 23–45.

Donaldson, S. I. (2007). *Program theory–driven evaluation science: Strategies and applications*. Erlbaum.

Education Sciences Reform Act of 2002 (Pub. L. No. 107-279). Retrieved June 29, 2022, from http://www.ed.gov/legislation/EdSciencesRef/

Ensminger, D. C., Frazier, E. W., Montrosse-Moorhead, B., & Linfield, K. J. (2021). How do we deepen our story reservoir by designing, developing, and writing instructional cases for teaching evaluation? *New Directions for Evaluation*, *172*, 85–102. https://doi.org/10.1002/ev.20484

Fuchs, D., Fuchs, L. S., Bahr, M. W., Fernstrom, P., & Stecker, P. M. (1990). Prereferral intervention: A prescriptive approach. *Exceptional Children*, *56*(6), 493–513.

Funnell, S. C., & Rogers, P. J. (2011). *Purposeful program theory: Effective use of theories of change and logic models*. John Wiley & Sons.

Gawande, A. (2014). *The Checklist Manifesto*. Penguin Books.

Gomaa, W. H. (2013). A survey of text similarity approaches. *International Journal of Computer Applications*, *68*, 6.

Hall, G. E., & Loucks, S. F. (1977). A developmental model for determining whether the treatment is actually implemented. *American Educational Research Journal*, *14*(3), 263–276.

Havelock, R. G. (1969). *A comparative study of the literature on the dissemination and utilization of scientific knowledge*. Center for Research on Utilization of Scientific Knowledge at the University of Michigan. https://files.eric.ed.gov/fulltext/ED029171.pdf

Hill, H. C. (2011). The nature and effects of middle school mathematics teacher learning experiences. *Teachers College Record*, *113*(1), 205–234.

Hill, H. C., & Erickson, A. (2019). Using implementation fidelity to aid in interpreting program impacts: A brief review. *Educational Researcher*, *48*(9), 590–598. https://doi.org/10.3102/0013189X19891436

House, E. R., Kerins, T., & Steele, J. M. (1972). A test of the research and development model of change. *Educational Administration Quarterly*, *8*(1), 1–14. https://doi.org/10.1177/0013131X7200800102

Hulleman, C. & Cordray, D. (2009). Moving from the lab to the field: The role of fidelity and achieved relative strength. *Journal of Research on Educational Effectiveness*, *2*, 88–110.

Institute of Education Sciences (2013). *FY 2013 Education Research Grants RFA*.

Institute of Education Sciences (2021). *Standards for Excellence in Educational Research (SEER) Principles*. https://ies.ed.gov/seer/index.asp , retrieved 6/23/22.

Institute of Education Sciences (2023). *FY 2024 Education Research Grants RFA*.

Jabeen, S. (2016). Do we really care about unintended outcomes? An analysis of evaluation theory and practice. *Evaluation and Program Planning*, *55*, 144–154. https://doi.org/10.1016/j.evalprogplan.2015.12.010

Jensen, E., Dale, M., Donnelly, P. J., Stone, C., Kelly, S., Godley, A., & D'Mello, S. K. (2020).

Toward automated feedback on teacher discourse to enhance teacher learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3313831.3376418

LaVelle, J. M. (2018). *2018 Directory of evaluator education programs in the United States. University of Minnesota Libraries Publishing*. https://conservancy.umn.edu/items/c14d5684-9030-4b86-9009-c11e0b095c6f.

Lewis, C., & Hurd, J. (2011). *Lesson study step by step: How teacher learning communities improve instruction*. Heinemann.

Lewis, C., & Tsuchida, I. (1997). Planned educational change in Japan: The case of elementary science instruction. *Journal of Educational Policy*, *12*(5), 313–331.

Lipsey, M. W., Crosse, S., Dunkle, J., Pollard, J., & Stobart, G. (1985). Evaluation: The state of the art and the sorry state of the science. *New Directions for Program Evaluation*, *27*, 7–28.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Meyers, C. V., & Brandt, W. C. (Eds.). (2014). *Implementation fidelity in education research: Designer and evaluator considerations*. Routledge.

Montrosse-Moorhead, B., Juskiewicz, K., & Li, E. Y. (2016). *Have we reached consensus on implementation fidelity in evaluation practice?* [Conference presentation] Annual meeting of the American Educational Research Association, Washington, DC, United States.

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

Munter, C., Wilhelm, A. G., Cobb, P., & Cordray, D. S. (2014). Assessing fidelity of implementation of an unprescribed, diagnostic mathematics intervention. *Journal of Research on Educational Effectiveness*, *7*(1), 83–113.

NIER (National Institute for Educational Policy Research) [Kokuritsu Kyouiku Seisaku Kenkyuujo]. (2011). *Kyouin no Shitsu no koujou ni kansuru chosa kenkyuu* [Report of Survey Research on Improvement of Teacher Quality]. Kokuritsu Kyouiku Seisaku Kenkyuujou.

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *The Journal of Behavioral Health Services and Research, 39*(4): 374–396.

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

Orwin, R. G., Sonnefeld, L. J., Cordray, D. S., Pion, G. M., & Perl, H. I. (1998). Constructing quantitative implementation scales from categorical services data: Examples from a multisite evaluation. *Evaluation Review*, *22*(2), 245–288.

Perry, R. R., & Lewis, C. C. (2009). What is successful adaptation of lesson study in the US? *Journal of Educational Change*, 1ß0, 365–391.

Rogers, E. M. (1995) *Diffusion of innovations*. Free Press.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Sage.

Scriven, M. (1973). Goal-free evaluation. In E. R. House (Ed.), *School evaluation: The politics and the process*. McCutchan Publishing.

Sechrest, L., & Redner, R. (1979). *Strength and integrity of treatments in evaluation studies: How well does it work?* National Criminal Justice Reference Service.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*(2), 5–24. https://doi.org/10.1111/j.1745-3992.1997.tb00585.x

Stigler, J. W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom.* Summit Books.

WALS. (2012). *World Association of Lesson Studies International Conference 2014: Programme and abstracts*. World Association of Lesson Studies. http://www.walsnet.org/2012/programme.html

Wang-Iverson, P., & Yoshida, M. (2005). *Building our understanding of lesson study*. Research for Better Schools.