Fidelity of Implementation: An Overlooked Yet Critical Construct to Establish Effectiveness of Evidence-Based Instructional Practices

Marilyne Stains* and Trisha Vickrey
Department of Chemistry, University of Nebraska–Lincoln, Lincoln, NE 68588

ABSTRACT

The discipline-based education research (DBER) community has been invested in the research and development of evidence-based instructional practices (EBIPs) for decades. Unfortunately, investigations of the impact of EBIPs on student outcomes typically do not characterize instructors' adherence to an EBIP, often assuming that implementation was as intended by developers. The validity of such findings is compromised, since positive or negative outcomes can be incorrectly attributed to an EBIP when other factors impacting implementation are often present. This methodological flaw can be overcome by developing measures to determine the fidelity of implementation (FOI) of an intervention, a construct extensively studied in other fields, such as healthcare. Unfortunately, few frameworks to measure FOI in educational settings exist, which likely contributes to a lack of FOI constructs in most impact studies of EBIPs in DBER. In this *Essay*, we leverage the FOI literature presented in other fields to propose an appropriate framework for FOI within the context of DBER. We describe how this framework enhances the validity of EBIP impact studies and provide methodological guidelines for how it should be integrated in such studies. Finally, we demonstrate the application of our framework to peer instruction, a commonly researched EBIP within the DBER community.

INTRODUCTION

One of the main goals of the discipline-based education research (DBER) community is to enhance how science, technology, engineering, and mathematics (STEM) are taught to college students. Decades of research and development activities to attain this goal have resulted in the characterization of evidence-based instructional practices (EBIPs). Misset and Foster probably described best what these practices are: "Broadly speaking, evidence-based practices consist of clearly described curricular interventions, programs, and instructional techniques with methodologically rigorous research bases supporting their effectiveness" (Missett and Foster, 2015, p. 97). Within the past couple of years, many initiatives have focused on the propagation of these practices to college STEM classrooms with the goal of increasing students' understanding and retention in STEM fields (American Association for the Advancement of Science, 2011; American Association of Universities, 2011; National Research Council [NRC], 2011, 2012; President's Council of Advisors on Science and Technology, 2012). However, the success of these propagation efforts relies on understanding how these practices can be broadly implemented with the same level of quality intended by the EBIPs' developers.

The research characterizing the impact of EBIPs on student outcomes (e.g., learning, retention, and affect) has typically been carried out using some form of experimental design: the practice is implemented by the designer(s) of the EBIP or a DBER-informed instructor, and the outcomes of the implementation of the EBIP are then compared with a control implementation. This control implementation can be

Deborah Allen, *Monitoring Editor*
Submitted March 7, 2016; Revised December 14, 2016; Accepted December 20, 2016
CBE Life Sci Educ March 1, 2017 16:rm1
DOI:10.1187/cbe.16-03-0113
*Address correspondence to: Marilyne Stains (mstains2@unl.edu).

carried out in different ways: either the designer of the EBIP or a DBER-informed instructor teaches a different section of the course in a traditional format; or the control implementation is the same course but taught in semesters or years before the EBIP implementation; or a traditional instructor teaches another section of the same course. Although these kind of experiments, also referred to as *efficacy studies*, are critical to establish the viability and efficacy of EBIPs, the outcomes identified in these studies may not be reflective of outcomes that would be observed in studies in real-world environments, also referred to as *effectiveness studies*, in which EBIPs are implemented by instructors "at-large." Indeed, some researchers recently conducted a large-scale study to characterize the relationship between levels of active-learning methods implemented by typical biology instructors and student learning gains (Andrews *et al.*, 2011). Their analyses revealed that active learning and learning gains were not related. The authors concluded that typical instructors lack the knowledge to implement active-learning teaching methods effectively. This conclusion is consistent with other DBER studies demonstrating that instructors often adapt EBIPs and unknowingly remove features critical to the efficacy of the EBIPs (Henderson and Dancy, 2009; Turpen and Finkelstein, 2009; Chase *et al.*, 2013; Daubenmire *et al.*, 2015).

Adaptations to EBIPs are inevitable in educational settings (Hall and Loucks, 1977, 1978). Factors such as class size, pressure of content coverage, and instructors' personal views on and prior experiences with teaching have been demonstrated to impact instructional decisions (Gess-Newsome *et al.*, 2003; Gess-Newsome, 2015; Henderson and Dancy, 2007; Andrews and Lemons, 2015; Lund and Stains, 2015). Unfortunately, few empirical studies have characterized elements of EBIPs' implementation that are critical to achieve expected outcomes (NRC, 2012). Exceptions include work by Linton on critical features of active learning (Linton *et al.*, 2014a,b) and research on peer instruction (Vickrey *et al.*, 2015). This dearth of knowledge of features critical to EBIPs' effectiveness not only results in uninformed and ineffective implementation of EBIPs but also calls into question the validity of efficacy and effectiveness research studies characterizing the impact of EBIPs. Without a detailed description of how an EBIP is implemented within a research study and the extent to which the control setting is similar to or different from the EBIP's implementation, the presence or absence of positive outcomes cannot reliably be attributed to the success or failure of the EBIP: any success observed may be due to factors other than the EBIP and lack of success may be due to improper implementation of the EBIP, not the EBIP itself (Lastica and O'Donnell, 2007). It is therefore essential to measure features of EBIPs' implementations within the control and treatment instructional environments to derive valid claims about the effectiveness of the EBIPs. These empirical investigations would enable the identification of the features of EBIPs that are critical to positive student outcomes and would inform instructors on appropriate adaptations.

The goals of this *Essay* are to provide the DBER community with both a framework and methodological approach for identifying the critical features of an EBIP and to enable future researchers to measure the extent to which these features are adhered to during implementation. Here, we present our framework and demonstrate its potential by applying it to a well-researched and disseminated EBIP, peer instruction (PI) (Mazur, 1997; Crouch and Mazur, 2001; Fagen *et al.*, 2002; Henderson and Dancy, 2009; Vickrey *et al.*, 2015).

OPENING THE INTERVENTION BLACK BOX: FIDELITY OF IMPLEMENTATION AS A FRAMEWORK TO EXPLAIN THE IMPACT OF AN EBIP

> Only by understanding and measuring whether an intervention has been implemented with fidelity can researchers and practitioners gain a better understanding of how and why an intervention works, and the extent to which outcomes can be improved. (Carroll *et al.*, 2007, p. 1)

Impacts of EBIPs (impact studies) have typically been established by assessing differences in outcomes between a treatment (i.e., implementation of an EBIP) and a control (business-as-usual) educational setting. This type of investigation helps answer the question "Does the EBIP work?" but does not capture why, how, and under what conditions the EBIP is impactful (Century *et al.*, 2010). The inner processes of the implementation of an EBIP and differences between these processes and those within the business-as-usual environment are not taken into account and are not related to the relative strengths of the measured outcomes. Moreover, comparisons of the impact of EBIPs across different treatment settings typically assume that the EBIPs are implemented as intended by the designers in all settings and that no other factors could contribute to observed outcomes. The intervention is thus treated as a "black box" (Century *et al.*, 2010). This lack of characterization of how EBIPs are implemented, factors that can moderate their implementation, and the relationship between the characteristics of the EBIPs' implementation, moderating factors, and outcomes of the implementation threaten the validity of the conclusions made regarding the effectiveness of the EBIPs. The U.S. Department of Education and the National Science Foundation have attempted to address this critical methodological problem by calling for the measurement of fidelity of implementation (FOI) when conducting impact studies and analyzing relationships between variations in FOI and intervention outcomes (Institute of Education Sciences, 2013).

FOI has been studied in various disciplines from health to manufacturing to science education. The variety of disciplines and cultures of study has resulted in different conceptualizations of FOI. Common to these different conceptualizations is the idea that certain essential features of the intervention must be measured in order to claim that the intervention has actually been implemented (Century *et al.*, 2010). Although these features have received various labels in the literature, the most common name used has been *critical components* (Century *et al.*, 2010). Century and colleagues provide an integrated definition of FOI with critical components that helps clarify and unify prior conceptualizations of FOI and its measurement: "We operationalized our FOI definition by rewording it as the extent to which the critical components of an intended program are present when that program is enacted" (Century *et al.*, 2010, p. 202). For the purposes of this *Essay*, we have slightly adapted this definition to fit our DBER context: Fidelity of implementation represents the extent to which the critical components of

an intended educational program, curriculum, or instructional practice are present when that program, curriculum, or practice is enacted.

From a research perspective, the measurement of FOI and its role in assessing the impact of an intervention enhance the validity of the research design and findings (Moncher and Prinz, 1991; Ruiz-Primo, 2006; Lastica and O'Donnell, 2007; O'Donnell, 2008). In particular, FOI helps establish internal validity (i.e., causal relationships between an intervention and outcomes are measured and other factors moderating intervention outcomes are controlled for) and construct validity (i.e., the intervention is implemented as intended) of an impact study by providing evidence for the extent to which the EBIP's implementation followed designers' recommendations. FOI also promotes the external validity of impact studies by enabling the identification of critical parameters to the success of the EBIP that are generalizable across settings.

On a practical level, FOI helps promote the successful propagation of an EBIP. Indeed, results of impact studies in which FOI has been measured include empirically established, detailed descriptions of how the EBIP is to be implemented and under what conditions in order to obtain desired outcomes. This information is critical to inform instructors about threatening and nonthreatening adaptations they can make to the EBIP; to support facilitators of professional development programs targeting the EBIP; and to help administrators, who are either contemplating the endorsement of an EBIP, promoting its use, or analyzing its impact on their campus, to make informed decisions.

FIDELITY OF IMPLEMENTATION FRAMEWORK

Although FOI has been extensively investigated in various intervention fields, including in K–12 STEM education (e.g., Penuel and Means, 2004; Lee *et al.*, 2009; Plass *et al.*, 2012; McNeill *et al.*, 2013) for more than a decade, its recognition within the DBER community has only grown over the past couple of years. The NRC report on the status, contributions, and future directions of DBER identified FOI-based studies as one of the necessary directions for future research on EBIPs: "However, much more research remains to be done to investigate how these pedagogies can best be implemented, how different student populations are affected, and how the fidelity of implementation—that is, the extent to which the experience as implemented follows the intended design—affects outcomes" (NRC, 2012, p. 126). Several DBER studies published in this journal in the past couple of years have addressed FOI in various fashions. For example, Tanner (2011) highlighted the need to develop a common language to describe what takes place when an EBIP is implemented in order to increase FOI; Eddy and Hogan (2014) studied conditions and populations for which course structure is most impactful and thus helped characterize factors that may mediate outcomes; Linton *et al.* (2014a,b) investigated essential features of active-learning teaching practices; and Corwin *et al.* (2015) developed a tool to measure the extent to which critical features of course-based undergraduate research experiences (CUREs) are being implemented. DBER researchers in various disciplines are increasingly including some measures or reports of FOI in their impact studies (e.g., Chase *et al.*, 2013; Drits-Esser *et al.*, 2014; Chan and Bauer, 2015) and exploring FOI for various EBIPs (e.g., Henderson and Dancy, 2009; Turpen and Finkelstein, 2009; Ebert-May *et al.*, 2011; Borrego *et al.*, 2013).

This increasing interest in FOI raises the need for the DBER community to identify a common framework to think about and measure FOI. This framework would help the community design impact studies with appropriate measures and employ analytical approaches that would enhance the validity of claims about the effectiveness of EBIPs. The FOI framework presented in this *Essay* draws primarily from the framework described in a recent article by Century *et al.* (2010). In what follows, we will describe the two main components of the FOI framework as it relates to DBER studies, and we later apply this FOI framework to PI.

Categories of Critical Components

Two main types of critical components have been measured in the FOI literature: structural and process components. Structural components relate to expected organizational features of the interventions (e.g., materials needed, frequency of use of certain activities); and process components relate to how the intervention is expected to be implemented, such as the expected behaviors of both instructors and students (Mowbray *et al.*, 2003; Century *et al.*, 2010). Prior FOI research indicates that quality investigations of FOI require the measurement of both structural and process components (Mowbray *et al.*, 2003; O'Donnell, 2008; Century *et al.*, 2010): "Distinctions should be made between measuring fidelity to the structural components of a curriculum intervention and fidelity to the processes that guide its design. […] Therefore, researchers should measure fidelity to both the structure and processes of an intervention, and relate both to outcomes" (O'Donnell, 2008, pp. 51, 52). Century *et al.* (2010) further divide the structural and process critical component categories into subcategories to better characterize the relationships between intervention outcomes and distinguishable yet complementary critical components within each of these broader critical component categories. These categories and subcategories are synthesized in Figure 1 and outlined below.

Structural Critical Components: Procedural. This subcategory of structural critical components characterizes the designer's intent about what the instructors should do (Century *et al.*, 2010). It includes descriptions of how the program, curriculum, or practice is intended to be implemented, with a focus on procedures and organizational features of the program, curriculum, or practice. Examples of potential critical components in this subcategory include the expected length of time of the intervention, order of instructional elements of the intervention, and nature of instructional materials.

Structural Critical Components: Educative. This subcategory of structural critical components describes the designers' expectations for the body of knowledge that instructors must possess in order to achieve high FOI of the program, curriculum, or practice (Century *et al.*, 2010). Examples include content knowledge, pedagogical knowledge, and knowledge of assessment.

Instructional Critical Components: Pedagogical. This subcategory of instructional critical components identifies the expectations for the instructor's behaviors and interactions with

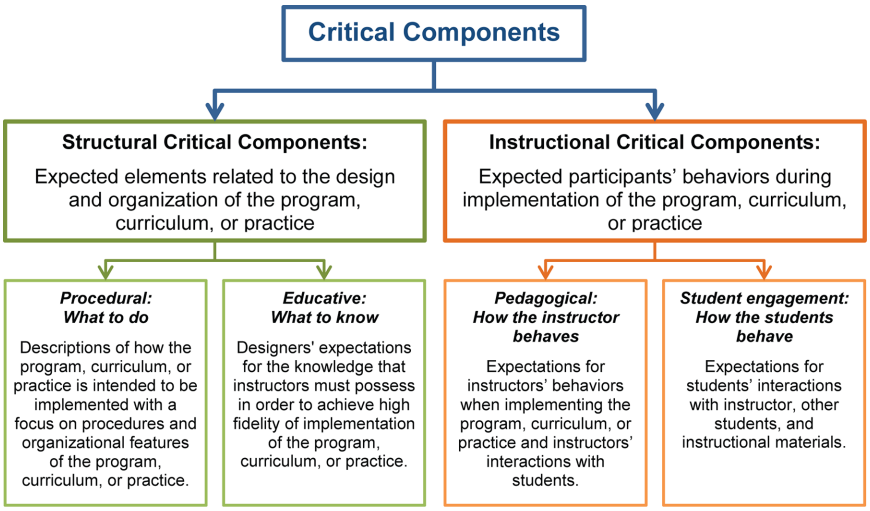

FIGURE 1. Categories and subcategories of critical components.

students when implementing the program, curriculum, or practice (Century *et al.*, 2010). Examples of possible critical components within this subcategory includes asking students to make predictions before watching the outcome of a demonstration, using empirical data during lecture, and facilitating a whole-class discussion.

Instructional Critical Components: Student Engagement. This subcategory of instructional critical components identifies the expectations for students' interactions with an instructor, other students, and instructional materials (Century *et al.*, 2010). Students asking questions, discussing their ideas in a small group, and developing strategies to solve problems are examples of potential critical components within this subcategory.

Moderating Variables

Implementation of an educational program, curriculum, or instructional strategy also depends on components outside the intervention. For example, characteristics of the participants (e.g., participation, enthusiasm, or teaching context of the participant) or delivery of the professional development introducing the intervention (e.g., enthusiasm or attitude of the workshop facilitator) may potentially mediate or moderate FOI even when appropriate structural and process critical components are present in an intervention (Dane and Schneider, 1998; Ruiz-Primo, 2006; Carroll *et al.*, 2007). To achieve the ultimate goal of relating FOI to outcomes, it is thus critical to identify potential moderators (Ruiz-Primo, 2006; Carroll *et al.*, 2007; Century *et al.*, 2010). Within the context of DBER, research has demonstrated that conceptions of teaching (e.g., attitudes, values, and goals of individual instructors), teaching context (e.g., departmental expectations, reward structure, time, class layout), and students' resistance are salient factors influencing the instructional practices of STEM faculty (Sunal *et al.*, 2001; Henderson and Dancy, 2007; Henderson *et al.*, 2011; Seidel and Tanner, 2013; Lund and Stains, 2015). Moderating variables within each of these categories should be characterized, measured, and integrated in the analysis of the relationship between FOI and intervention outcomes in order to comprehensively characterize the reasons and context behind the success of an intervention (Century *et al.*, 2010; Institute of Education Sciences, 2013). Furthermore, these analyses should be situated specifically within the STEM teaching context.

CHARACTERIZING FIDELITY OF IMPLEMENTATION

To articulate a methodological approach to measure FOI that is consistent with the above framework, we leverage several literature reviews (Mowbray *et al.*, 2003; Ruiz-Primo, 2006; O'Donnell, 2008; Century *et al.*, 2010; Missett and Foster, 2015) focused on the methods used to characterize FOI. In the next sections, we will describe an approach to identify, measure, and validate critical components as well as strategies used to derive a final measure of FOI that can be used within DBER.

Methods to Identify Potential Critical Components

The first step in developing a measure for FOI is the identification of potential critical components. Mowbray *et al.* (2003) reviewed studies on FOI conducted from 1995 to 2003 in mental health, health, substance abuse treatment, education, and social services fields, and identified three main approaches to the characterization of critical components: 1) leveraging empirical studies, 2) consulting experts, and 3) conducting qualitative research. Leveraging empirical studies consists of analyzing results of efficacy and effectiveness studies. Consulting experts includes collecting and analyzing surveys and interviews with designers of the intervention, conducting a literature review on the intervention, and analyzing the materials used to disseminate and propagate the intervention to potential users. Finally, conducting qualitative research consists of questioning users of the intervention (instructors and students) about their perceptions of the strengths and weaknesses of the intervention, conducting observations during training of instructors, and analyzing diverse types of implementation. Ideally, critical components are characterized through a mixture of these three approaches to overcome limitations associated with each method (Ruiz-Primo, 2006). Moreover, researchers should follow an iterative process, consisting of consulting intervention designers on critical components emerging from the data, refining these components accordingly, and identifying *suspected* productive, neutral, and unproductive adaptations of each component (Mills and Ragan, 2000; Ruiz-Primo, 2006; Century *et al.*, 2010). Importantly, the critical components identified and chosen should be measurable and at a level of granularity that does not impede the feasibility and meaningfulness of data collection and analysis (Ruiz-Primo, 2006).

Methods to Measure and Validate Critical Components

Once potential critical components and their suspected different levels of adaptation have been identified, a strategy and tools to characterize the presence of these components during implementation should be developed. It is understood that

a multimethod approach (e.g., observations conducted by experts, self-reports from instructors and students, interviews, and intervention artifacts) involving various stakeholders (e.g., designers of the intervention, other experts within the field, instructors, and students) is preferred (Mowbray *et al.*, 2003; O'Donnell, 2008; Nelson *et al.*, 2012; Missett and Foster, 2015). This approach will enhance the validity and reliability of the final evaluation of the level of FOI.

However, before using these tools to determine a method to calculate the overall level of FOI, it is critical to investigate the reliability and validity of the identified critical components (Mowbray *et al.*, 2003; Nelson *et al.*, 2012; Institute of Education Sciences, 2013). Reliability and validity should be established by employing at least one of the following methods (for further reading, see Mowbray *et al.*, 2003):

1. Content validity is typically established by involving designers of the intervention in the determination of the critical components and their adaptations.
2. Known-groups validity is established by characterizing whether differences on measures of critical components are observed between interventions that are known for implementing these components differently.
3. Convergent validity is established by identifying the level of agreement on the critical components across two or more data sources.
4. Internal structure is established by conducting statistical analyses such as factor analysis or cluster analysis to identify the coherence across different measures intended to characterize the extent of implementation of a specific critical component.
5. Interrater reliability is established by calculating the level of agreement (i.e., kappa, intraclass correlations, percent agreement) between two or more users on critical components' measures such as rubrics used to analyze implementation artifacts or codes used to analyze interviews.
6. Internal consistency reliability is evaluated by calculating Cronbach's alpha for certain types of measures of critical components, such as surveys or questionnaires.

Methods to Determine Implementation Types

Researchers have taken different approaches to measuring the level of FOI (Century *et al.*, 2010). Most involve calculating the proportion of critical components implemented, treating all components with equal importance, and delivering one number to express the level of FOI (e.g., Gresham, 1989; Mills and Ragan, 2000; Balfanz *et al.*, 2006). Others have used the proportion of users implementing a specific number of critical components (e.g., Borrego *et al.*, 2013) or time spent by all users on each component (e.g., Borrego *et al.*, 2013; Barker *et al.*, 2014). Unfortunately, these approaches do not account for variations in the implementation of each critical component or that some critical components might be more important than others. Therefore, an approach is needed that identifies 1) productive and unproductive adaptations (e.g., adaptations to critical components that lead or do not lead to desired intervention outcomes), 2) the necessary combinations of critical components required to achieve desired outcomes, and 3) intervention components that require further attention during dissemination and propagation of the intervention. This approach to measuring the level of FOI better elucidates the role of each potential critical component in the success of the intervention as well as the extent to which critical components can be adapted without compromising the effectiveness of the intervention (Century *et al.*, 2010).

Thus, to more comprehensively determine the level of FOI, we recommend using "implementation type," which has been defined by Century *et al.* as "a particular combination of critical components enacted to particular degrees" (Century *et al.*, 2010, p. 213). Implementation type is derived from the four scores obtained in each of the critical component subcategories (i.e., structural–procedural, structural–educative, instructional–pedagogical, and instructional–student engagement). The score for each critical component subcategory is based on the presence of and types of adaptation made to the critical components within the subcategory. The implementation that yields the highest level of FOI is identified by investigating the relationship between implementation types and desired outcomes of the intervention. For example, one may find that an implementation type that has high scores on structural–educative and instructional–pedagogical critical components and low scores on structural–procedural and instructional–student engagement critical components leads to better student outcomes than an implementation type with high scores on both instructional critical component subcategories and low scores on both structural critical component subcategories. The former implementation type will then represent a higher FOI than the latter.

APPLICATION OF THE FIDELITY OF IMPLEMENTATION FRAMEWORK TO ESTABLISH EFFECTIVENESS OF EBIPS

The FOI framework presented in this *Essay* enables DBER researchers to move beyond an all-or-nothing conclusion regarding the success of an intervention. In particular, it permits the characterization of the extent to which the intervention was effective and the identification of causes behind the success or lack thereof of the intervention. In this section, we describe a methodological approach to determine the effectiveness of an EBIP that leverages both critical components and level of FOI described in the above framework. Our approach is based on prior work in FOI (Mowbray *et al.*, 2003; O'Donnell, 2008; Century *et al.*, 2010; Missett and Foster, 2015) and national standards for impact studies in education (Institute of Education Sciences, 2013), which indicate that establishing the impact of an intervention requires both efficacy and effectiveness studies as well as FOI measurements (i.e., critical components and implementation type) in both treatment and control/comparison groups. We summarize our approach for using efficacy and effectiveness studies in Figure 2 and discuss each in further detail below.

Efficacy studies typically build on well-designed pilot studies that demonstrate promising outcomes of an intervention. They are conducted in an ideal setting, one example being an intervention implemented by highly trained instructors (O'Donnell, 2008; Institute of Education Sciences, 2013). Efficacy studies help identify critical components and implementation types that are linked to positive outcomes and the range of adaptations to each critical component and its impact on the effectiveness of the intervention (Century *et al.*, 2010). In particular, efficacy studies designed to determine the impact

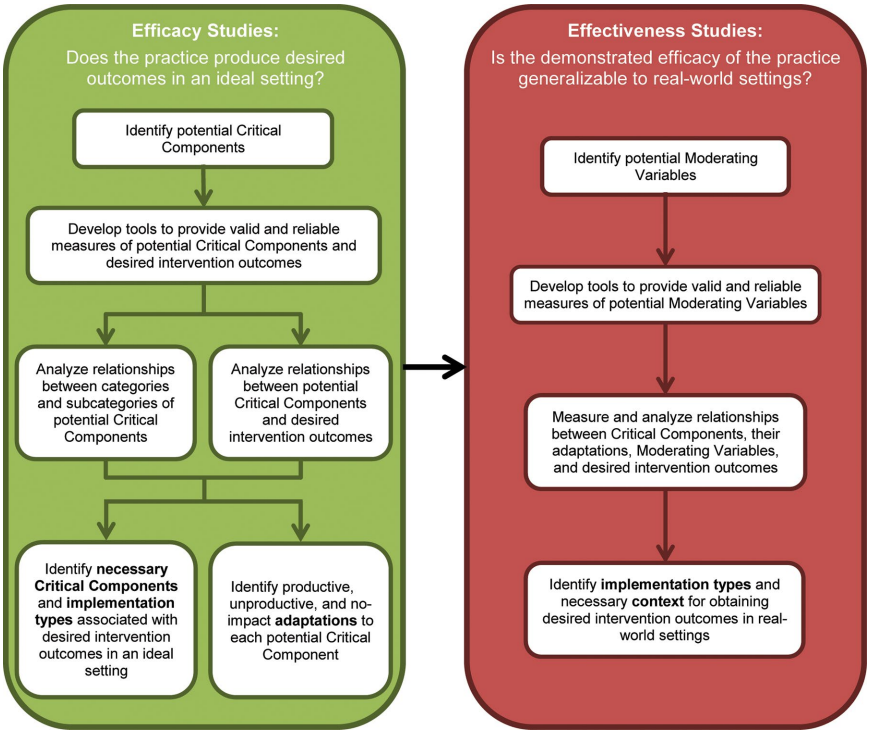

FIGURE 2. Process to establish the effectiveness of an EBIP using the fidelity of implementation framework.

of implementing a strategy in the presence, absence, or adaptation of a critical component (e.g., peer discussion during PI) will elucidate the importance of that component. We propose that the following steps be used to design an efficacy study (Figure 2):

1. Identify potential critical components: see *Methods to Identify Potential Critical Components* section;
2. Develop tools to measure the presence of these potential critical components: see *Methods to Measure and Validate Critical Components*;
3. Develop tools (e.g., a concept inventory) to provide valid and reliable data about intervention outcomes (for a detailed methodological description of the process required to develop a concept inventory, see Adams and Wieman (2011);
4. Implement the intervention in both a treatment (e.g., with a potential critical component) and control setting (e.g., without the potential critical component), and collect data in each setting using the measurement tools developed for the potential critical components and intervention outcomes;
5. Analyze the correlations between categories and subcategories of potential critical components to identify presence or absence of expected relationships (Century *et al.*, 2010); this analysis can inform required adaptations and provide more in-depth understanding of the nature of the relationship between the implementation of the intervention and its outcomes;
6. Analyze the relationship between potential critical components and the desired outcomes for the intervention; this analysis will inform the selection of the final list of critical components.

Triangulate findings to characterize the nature and level of threat to intervention outcomes of adaptations identified for each critical component, and develop a final list of necessary critical components and implementation types that yield highest intervention outcomes (see *Methods to Determine Implementation Types*).

This process will be iterative (Century *et al.*, 2010): intervention, critical components, adaptations, and implementation types will be adjusted and revised as results of efficacy studies emerge. Once enough evidence supporting the positive relationships between intervention outcomes and critical components and implementation types exist, effectiveness studies can be conducted (O'Donnell, 2008; Institute of Education Sciences, 2013). Effectiveness studies are necessary to demonstrate the generalizability of the intervention outcomes obtained in ideal contexts to real-world contexts (O'Donnell, 2008; Institute of Education Sciences, 2013). We propose an effectiveness study be carried out using the following steps (Figure 2):

1. Identify potential moderating variables: see *Moderating Variables* section; identification methods can include a literature review; interviews with and/or surveys from instructors, developers, and students; and observations of implementation in different instructional settings;
2. Develop tools to measure the presence of these moderating variables;
3. Implement the intervention and collect data in treatment and control settings using the measurement tools for the critical components, intervention outcomes, and moderating variables (Institute of Education Sciences, 2013);
4. Analyze relationships between critical components, their adaptations, moderating variables, and desired intervention outcomes (Mowbray *et al.*, 2003; Missett and Foster, 2015);
5. Derive from the findings implementation types (see *Methods to Determine Implementation Types*) and context that lead to desired intervention outcomes in real-world settings.

Outcomes of these studies will include rigorous empirical evidence supporting the effectiveness of the EBIP and empirical understanding of the specificity of the implementation that leads to positive outcomes. These results can also be leveraged by professional development program facilitators to more effectively propagate the practice.

APPLICATION OF THE FIDELITY OF IMPLEMENTATION FRAMEWORK TO PEER INSTRUCTION

To exemplify how the FOI framework described in this *Essay* can be applied to the study of EBIPs, we applied it to the EBIP known as peer instruction (PI). This practice consists of

interspersing challenging questions throughout a lecture; students are first asked to vote on each question individually, using either a personal response system (e.g., clicker) or flash cards; on the basis of the outcome of this vote, the instructor either provides a short explanation or asks students to discuss with one another potential answers to the question; in the latter case, a revote is conducted after a few minutes of peer discussion. PI is probably one of the most well-researched EBIPs in DBER, with many efficacy studies in a variety of STEM fields having been conducted. More detailed information about this practice and previous empirical investigations can be found in a review *Essay* published in this journal (Vickrey *et al.*, 2015).

Identification of Potential Critical Components

Potential critical components for the implementation of PI were identified using the three methods described earlier in the *Methods to Identify Potential Critical Components* section: 1) leveraging empirical studies, 2) consulting experts, and 3) conducting qualitative research. Regarding the first method, we primarily relied on an earlier *Essay* that we published in this journal. For that *Essay*, we conducted a comprehensive review of the literature that provided empirical support to the impact or lack thereof of each step involved in the implementation of PI (Vickrey *et al.*, 2015). Using the same selection criteria detailed in this previous work, we also reviewed subsequent research on PI that followed the publication of our *Essay*. The findings from these efficacy studies were then used to identify and categorize critical components. For example, several studies indicate that posing more difficult questions attenuates learning outcomes. Thus, question difficulty was identified as a structural–procedural critical component of PI.

For the second and third methods, we leveraged the findings of prior studies in which researchers had already conducted such investigations. In particular, we leveraged a study conducted by engineering and physics education researchers in which they identified a certain set of critical components for PI based on their understanding of the literature and their own expertise with this practice (Borrego *et al.*, 2013). We also used studies conducted by physics education researchers who interviewed developers and expert users of PI (Turpen *et al.*, 2010, 2016). Finally, our experience developing and implementing a professional development program on PI and our knowledge of the DBER literature related to PI helped populate the structural–educative category; we then compared and matched our list of critical components in this subcategory with a list provided by experts in PI (American Association of Physics Teachers, 2011).

The outcome of these processes is presented in Table 1, in which we provide the list of potential critical components to the implementation of PI and a short description of each component. The names of some of these critical components come directly from wording provided by experts (American Association of Physics Teachers, 2011), while others were explored in our literature review (Vickrey *et al.*, 2015).

Variations existed in the level of empirical evidence supporting the critical components identified. Critical components supported by several, rigorous empirical studies should be included in the final list of critical components to be investigated in effectiveness studies, while those supported by few or less rigorous empirical studies should be further investigated through efficacy studies. We represent the level of empirical evidence supporting each critical component using a scale that uses the number of investigations conducted, different types of methods used (e.g., qualitative and quantitative or mixed methods), and unique populations studied (e.g., courses in different disciplines, institutions, or course levels) in these investigations (see Table 2). This scale was previously used in an analysis of the literature on the diffusion of innovations in health-service delivery and organization (Greenhalgh *et al.*, 2004) and was itself adapted from criteria developed by the World Health Organization Evidence Network (Øvretveit, 2003). Moreover, similar approaches have been used to evaluate the empirical evidence on CUREs (Corwin *et al.*, 2015). We ranked the level of empirical evidence for each critical component from one to four with four representing the highest level (e.g., largest number of studies, methods used, and different populations) and one representing the lowest level of evidence. Table 1 includes this ranking for each critical component.

Identification of Moderating Variables

Few studies have attempted to characterize moderating variables to the implementation of PI and their influences (positive or negative) on the outcomes of PI implementation. As previously described, moderating variables are likely to be unique to a particular EBIP and teaching context; it is important to explore them for PI specifically. Thus, we leveraged artifacts collected during a professional development program on PI and a qualitative study conducted by physics education researchers (Turpen *et al.*, 2016) to identify the moderating variables presented in Table 3. Details about each source of moderating variables are provided next.

First, we used artifacts produced by participants from two cohorts (Fall 2014 and Fall 2015) of STEM faculty ($N = 24$) enrolled in a semester-long professional development program on PI facilitated by the authors. The workshop consisted of eight sessions, each 1.5 hours in length, and was designed to explicitly address the structural and instructional critical components described previously. During sessions 6 and 7, faculty practiced implementation of PI using a self-authored conceptual question and received feedback about FOI from peers and the authors. Before session 8, participants engaged in a short (three- to six-sentence) reflective writing activity asking them to 1) discuss how their experience with PI as both an instructor and student would inform their future implementation of PI, and 2) identify and discuss any potential barriers to the implementation of PI. Participant reflections revealed that there were individual, situational, and student-related perceptions influencing their use of critical components. For example, participants expressed increased student engagement as an affordance of PI; they also considered potential student resistance or a lack of time inside or outside of class as barriers to the implementation of critical components, such as developing conceptually challenging multiple-choice questions. One participant illustratively wrote, "Designing good questions is still the most salient [barrier], in my view. These have to be good, and conceptual, because they do take time in [class] to implement. They offer the opportunity to refocus students from passive observers to active learners, but need to be sufficiently engaging to promote productive student discussion and buy-in in the process." Although these artifacts have allowed us to identify some likely moderators for FOI, faculty self-selected to enroll in

TABLE 1. Critical components of the FOI framework for PI

Category	Subcategory	Critical component	Description	Level of empirical evidence
Structural	Procedural	Question difficulty	Challenging multiple-choice questions are being asked (fewer than two-thirds of the students chose the correct answer)	4
		Low-stakes grading	Points are primarily awarded for answering the question (participation)	3
		Questions interspersed	Questions are interspersed throughout the lecture	1
		Prior knowledge-based questions	Questions are based on knowledge of students' prior knowledge	1
	Educative	Constructivism	Students learn by constructing new knowledge based on their prior knowledge	1
		Collaborative learning	Students learn by working with others on a common goal	1
		Prior knowledge	Students' prior knowledge can positively or negatively interact with learning new knowledge	1
		Conceptual understanding	Students achieve deeper level of understanding when instruction is focused on concepts rather than memorization	1
		Verbalizing thinking	Students learn by providing verbal or written explanations of their thinking and understanding	1
		Formative assessments	Formative assessments support learning by providing feedback to both instructor and students	1
Instructional	Pedagogical	Reasoning-focused explanations	Explanations following the final vote are focused on the reasoning that led to the answer	4
		Decision to engage peer discussion	Decision to request students to discuss their answer or to move on with lecture is determined by the proportion of students who initially answer the question correctly	3
		Cuing reasoning	The instructor encourages students to focus on describing their reasoning during peer discussion	2
		Whole-class discussion	The instructor facilitates whole-class discussion following the final vote	2
		Incorrect answers	The instructor explains incorrect answers	2
		Moving during voting	The instructor walks around the classroom during voting, observing students	1
		Gaining students buy-in	The instructor explains the research supporting PI and their reasons to use it in the course	1
		Facilitating discussions	The instructor encourages students who are not engaged during peer discussion to talk to each other	1
		Listening to students	The instructor listens to students' conversations during peer discussion	1
		Use of histogram	The instructor only shows the histogram after the first vote if most students chose the correct answer, otherwise the instructor waits until the end of the second vote to show the histograms	1
		Closing vote	The instructor starts the countdown to finish voting once ~80% of the students have responded	1
	Student engagement	Peer discussion on second vote	Students discuss their answers in groups of two or more following an overall failed first vote by the class	4
		Individual voting on first vote	Students think about the question individually during the first vote	2

the PI workshop and, as a result, may be influenced by different moderators than faculty exposed to PI in other contexts, such as informal discussions with colleagues and reading.

We also leveraged a recent qualitative study conducted by physics education researchers (Turpen *et al.*, 2016). In this study, the authors interviewed self-reported users of PI ($N = 35$), who were exposed to PI in a variety of contexts, about their perceptions of barriers and affordances to the implementation of PI. The authors then classified interviewees based on their self-reported descriptions of their implementations of PI. Level of FOI of PI was characterized in this study from the proportion of nine essential features interviewees reported implementing: instructor adapts to student responses, uses low-stakes grading, gives time for individual voting, asks conceptual questions, targets known student difficulties, uses multiple-choice questions, intersperses questions, asks students to discuss questions with peers, and instructs students to revote after discussion. Interestingly, the prevalence of perceived affordances and barriers appeared to differ based on the extent to which they implemented all nine features. For example, out of the instructors using seven or more

TABLE 2. Definition of the scale used to characterize the level of empirical evidence for each critical component

Level of empirical evidence	Qualifier	Criterion
4	Strong evidence	Two or more studies on PI using two different research methods (e.g., qualitative and quantitative or mixed methods) on at least two different populations
3	Moderate evidence	Two or more studies on PI using the same research method (e.g., qualitative or quantitative) on at least two different populations
2	Limited evidence	One study on PI with a critical component as a research question AND/OR one or more studies on PI with indirect evidence in support of a critical component
1	No evidence established	No PI-related evidence but evidence in DBER or other educational research fields OR indeterminate evidence from PI research (e.g., two studies with opposing results)

key features of PI (high fidelity; $N = 7$), 71% perceived that PI improved student participation (an affordance of PI), compared with only 11% of instructors using one to six key features (mixed fidelity; $N = 18$), and 0% of nonusers ($N = 10$).

The affordances and barriers identified from this study and our workshop artifacts are consistent with previous research associating conceptions of teaching, teaching context, and student beliefs with instructional decision making (Gess-Newsome *et al.*, 2003; Gess-Newsome, 2015; Henderson and Dancy, 2007; Lund and Stains, 2015), but are more specific to the implementation of PI. We classified them into these three categories in Table 3. We propose that these affordances and barriers are tangible moderating variables for FOI of PI and should be investigated further.

Measures of Critical Components and Moderating Variables

The identification of critical components and moderating variables naturally led to a search for valid and reliable measures of these constructs. Unfortunately, we could find only one tool that aligns with the measure of some of the potential critical components identified in the FOI framework for PI: the Classroom Observation Protocol (Turpen and Finkelstein, 2009). This protocol measures some of the structural–procedural critical components (e.g., difficult questions are being asked and questions are interspersed) and several of the Instructional critical components (e.g., instructor walks around classroom, incorrect answers are discussed, group vs. individual voting). We found additional tools that could help measure the presence of moderating variables (Table 4). However, most of these tools were not specifically designed for the context of PI implementation and would need to be adapted when used to understand how these variables moderate FOI and student outcomes during PI implementation. Therefore, future work on FOI of PI should focus on designing and validating tools to measure the potential critical components and moderating variables we have identified.

SUMMARY

In this *Essay*, we argue for the need to measure FOI when characterizing the impact of EBIPs. This characterization is critical to provide validity to the findings of such studies and has been critically missing in DBER studies. We describe a DBER-specific framework for the measure of FOI that is based on prior empirical work on FOI in various fields. We hope that this framework will

TABLE 3. Moderating variables to the implementation of PI[a]

Type	Affordances	Barriers
Conceptions of teaching	• Dissatisfaction with traditional lecture • Encourages student engagement • Easy to incorporate into existing paradigm • Intuitively value PI • Evidence of effectiveness from personal experience or published data • Provides feedback • Students learn by working together • Promotes deep learning	• Requires too much time and energy • Satisfaction with current practice • Poor fit with personality • Intuitive disbelief in effectiveness of PI • Preference for other types of in-class assessments (e.g., open-ended questions)
Teaching context	• Departmental support or encouragement	• Class size (either too large or too small) • Classroom layout • External requirements for content coverage • Lack of resources to educate themselves about PI • Difficulty finding good questions
Student factors	• Buy-in	• Resistance • Students lack necessary knowledge and skills to engage appropriately

[a]Only affordances and barriers that were reported by at least a quarter of the interviewees in the Turpen *et al.* study (2016) as well as in the workshop artifacts are included in this table.

TABLE 4. Existing resources to measure moderating variables

Type of moderating variables	Measurement tools
Conceptions of teaching	Survey instruments
	Postsecondary Instructional Practices Survey (PIPS; Walter *et al.*, 2016)
	Teaching Practices Inventory (TPI; Wieman and Gilbert, 2014)
	Approaches to Teaching Inventory (ATI; Trigwell *et al.*, 2005)
	Pedagogical Discontentment Scale (Southerland *et al.*, 2012)
	Interview protocols
	Teacher Beliefs Interview (Luft and Roehrig, 2007)
Teaching context	Teaching Practices Inventory (TPI; Wieman and Gilbert, 2014)
	Survey of Climate for Instructional Improvement (SCII; Walter *et al.*, 2014)

standardize the studies on the impact of EBIPs within the DBER community. In turn, this will facilitate the dissemination and propagation of EBIPs by strengthening the validity of the findings and providing detailed and empirically tested descriptions of how EBIPs should be implemented to achieve desired outcomes.

The application of the FOI framework to PI and the search for measures of fidelity for the implementation of PI highlighted to us how extensive the gap is in this research arena. Indeed, even though PI is arguably one of the most extensively studied EBIPs in DBER, we struggled to find strong evidence supporting most potential critical components; tools to measure these critical components; and studies that investigate the relationships among the implementation of critical components, moderating variables, and outcomes of PI implementation. It is our hope that this *Essay* will inspire researchers to conduct efficacy and effectiveness studies to characterize the necessary critical components, moderating variables, and implementation types that lead to high FOI for PI. We also hope that this *Essay* will contribute to the growth of FOI-based investigations of the effectiveness of EBIPs.

Finally, while our *Essay* focused on carefully designing tools for measuring instructors' FOI of PI, we recognize that measuring student outcomes, and ultimately relating them to instructors' FOI, also requires careful design and consideration of student characteristics. Indeed, previous work by Theobald and Freeman (2014) in this journal demonstrates the importance of accounting for differences in student characteristics when assessing the outcome of an instructional intervention. Moreover, student characteristics such as self-efficacy and gender have been implicated as important predictors of student behavior (e.g., response switching) and performance during PI (Miller *et al.*, 2015). Additional research is needed to identify student characteristics that impact performance with this instructional strategy. Once predictive characteristics are better understood, this *educative* knowledge can be included in the FOI framework as a critical component and be accounted for during analyses.

ACKNOWLEDGMENTS

This material is based on work supported by the National Science Foundation (grant nos. DUE-1256003 and DUE-1347814).

REFERENCES

Adams WK, Wieman CE (2011). Development and validation of instruments to measure learning of expert-like thinking. Int J Sci Educ 33, 1289–1312.

American Association for the Advancement of Science (2011). Vision and Change in Undergraduate Biology Education: A Call to Action, Washington, DC.

American Association of Physics Teachers (2011). Peer instruction—what makes it work? In: PER User's Guide. www.compadre.org/perug/guides/Section.cfm?G=Peer_InstructionS=Why (accessed 28 February 2016).

American Association of Universities (2011). Undergraduate STEM Education Initiative. www.aau.edu/policy/article.aspx?id=12588 (accessed 30 December 2014).

Andrews TC, Lemons PP (2015). It's personal: biology instructors prioritize personal evidence over empirical evidence in teaching decisions. CBE Life Sci Educ 14, ar7 .

Andrews TM, Leonard MJ, Colgrove CA, Kalinowski ST (2011). Active learning not associated with student learning in a random sample of college biology courses. CBE Life Sci Educ 10, 394–405.

Balfanz R, Mac Iver DJ, Byrnes V (2006). The implementation and impact of evidence-based mathematics reforms in high-poverty middle schools: a multi-site, multi-year study. J Res Math Educ 37, 33–64.

Barker BS, Nugent G, Grandgenett NF (2014). Examining fidelity of program implementation in a STEM-oriented out-of-school setting. Int J Tech Des Educ 24, 39–52.

Borrego M, Cutler S, Prince M, Henderson C, Froyd JE (2013). Fidelity of implementation of research-based instructional strategies (RBIS) in engineering science courses. J Eng Educ 102, 394–425.

Carroll C, Patterson M, Wood S, Booth A, Rick J, Balain S (2007). A conceptual framework for implementation fidelity. Implement Sci 2, 40.

Century J, Rudnick M, Freeman C (2010). A framework for measuring fidelity of implementation: a foundation for shared language and accumulation of knowledge. Am J Eval 31, 199–218.

Chan JY, Bauer CF (2015). Effect of peer-led team learning (PLTL) on student achievement, attitude, and self-concept in college general chemistry in randomized and quasi experimental designs. J Res Sci Teach 52, 319–346.

Chase A, Pakhira D, Stains M (2013). Implementing process-oriented, guided-inquiry learning for the first time: adaptations and short-term impacts on students' attitude and performance. J Chem Educ 90, 409–416.

Corwin LA, Runyon C, Robinson A, Dolan EL (2015). The Laboratory Course Assessment survey: a tool to measure three dimensions of research-course design. CBE Life Sci Educ 14, ar37.

Crouch CH, Mazur E (2001). Peer instruction: ten years of experience and results. Am J Phys 69, 970–977.

Dane AV, Schneider BH (1998). Program integrity in primary and early secondary prevention: are implementation effects out of control? Clin Psychol Rev 18, 23–45.

Daubenmire PL, Bunce DM, Draus C, Frazier M, Gessell A, van Opstal MT (2015). During POGIL implementation the professor still makes a difference. J Coll Sci Teach 44, 72–81.

Drits-Esser D, Bass KM, Stark LA (2014). Using small-scale randomized controlled trials to evaluate the efficacy of new curricular materials. CBE Life Sci Educ 13, 593–601.

Ebert-May D, Derting TL, Hodder J, Momsen JL, Long TM, Jardeleza SE (2011). What we say is not what we do: effective evaluation of faculty professional development programs. BioScience 61, 550–558.

Eddy SL, Hogan KA (2014). Getting under the hood: how and for whom does increasing course structure work? CBE Life Sci Educ 13, 453–468.

Fagen AP, Crouch CH, Mazur E (2002). Peer instruction: results from a range of classrooms. Phys Teach 40, 206–209.

Gess-Newsome J (2015). A model of teacher professional knowledge and skill including PCK: results of the thinking from the PCK summit. In: Re-examining Pedagogical Content Knowledge in Science Education, ed. A Berry, P Friedrichsen, and J Loughran, New York: Routledge.

Gess-Newsome J, Southerland SA, Johnston A, Woodbury S (2003). Educational reform, personal practical theories, and dissatisfaction: the anatomy of change in college science teaching. Am Educ Res J 40, 731–767.

Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O (2004). Diffusion of innovations in service organizations: systematic review and recommendations. Milbank Q 82, 581–629.

Gresham FM (1989). Assessment of treatment integrity in school consultation and prereferral intervention. School Psych Rev 18, 37–50.

Hall GE, Loucks SF (1977). A developmental model for determining whether the treatment is actually implemented. Am Educ Res J 14, 263–276.

Hall GE, Loucks SF (1978). Innovation Configurations: Analyzing the Adaptations of Innovations, Austin: University of Texas, Research and Development Center for Teacher Education.

Henderson C, Beach A, Finkelstein N (2011). Facilitating change in undergraduate STEM instructional practices: an analytic review of the literature. J Res Sci Teach 48, 952–984.

Henderson C, Dancy MH (2007). Barriers to the use of research-based instructional strategies: the influence of both individual and situational characteristics. Phys Rev Spec Top Phys Educ Res 3, 020102.

Henderson C, Dancy MH (2009). Impact of physics education research on the teaching of introductory quantitative physics in the United States. Phys Rev Spec Top Phys Educ Res 5, 020107.

Institute of Education Sciences (2013). Common Guidelines for Education Research and Development: A Report from the Institute of Education Sciences and the National Science Foundation, Washington, DC: U.S. Department of Education.

Lastica J, O'Donnell C (2007). Considering the role of fidelity of implementation in science education research: Fidelity as teacher and student adherence to structure. Paper presented at the Annual Meeting of the American Educational Research Association, held 15–18 April 2007, in Chicago, IL.

Lee O, Penfield R, Maerten-Rivera J (2009). Effects of fidelity of implementation on science achievement gains among English language learners. J Res Sci Teach 46, 836–859.

Linton DL, Farmer JK, Peterson E (2014a). Is peer interaction necessary for optimal active learning?CBE Life Sci Educ 13, 243–252.

Linton DL, Pangle WM, Wyatt KH, Powell KN, Sherwood RE (2014b). Identifying key features of effective active learning: the effects of writing and peer discussion. CBE Life Sci Educ 13, 469–477.

Luft JA, Roehrig GH (2007). Capturing science teachers' epistemological beliefs: the development of the teacher beliefs interview. Electronic J Sci Educ 11(2), 38–63.

Lund TJ, Stains M (2015). The importance of context: an exploration of factors influencing the adoption of student-centered teaching among chemistry, biology, and physics faculty. Intl J STEM Educ 2(1), 1–21.

Mazur E (1997). Peer Instruction, Upper Saddle River, NJ: Prentice Hall.

McNeill KL, Pimentel DS, Strauss EG (2013). The impact of high school science teachers' beliefs, curricular enactments and experience on student learning during an inquiry-based urban ecology curriculum. Int J Sci Educ 35, 2608–2644.

Miller K, Schell J, Ho A, Lukoff B, Mazur E (2015). Response switching and self-efficacy in peer instruction classrooms. Phys Rev Spec Top Phys Educ Res 11, 010104.

Mills SC, Ragan TJ (2000). A tool for analyzing implementation fidelity of an integrated learning system. Educ Tech Res Dev 48(4), 21–41.

Missett TC, Foster LH (2015). Searching for evidence-based practice: a survey of empirical studies on curricular interventions measuring and reporting fidelity of implementation published during 2004–2013. J Adv Acad 26, 96–111.

Moncher FJ, Prinz RJ (1991). Treatment fidelity in outcome studies. Clin Psychol Rev 11, 247–266.

Mowbray CT, Holter MC, Teague GB, Bybee D (2003). Fidelity criteria: development, measurement, and validation. Am J Eval 24, 315–340.

National Research Council (NRC) (2011). Promising Practices in Undergraduate Science, Technology, Engineering, and Mathematics Education: Summary of Two Workshops, Washington, DC: National Academies Press.

NRC (2012). Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering, Washington, DC: National Academies Press.

Nelson MC, Cordray DS, Hulleman CS, Darrow CL, Sommer EC (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. J Behav Health Serv Res 39, 374–396.

O'Donnell CL (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. Rev Educ Res 78, 33–84.

Øvretveit J (2003). Reviewing Medical Management Research for Decision-Makers: Methodological Issues in Carrying Out Systematic Reviews of Medical Management Research, Medical Management Centre Internal Discussion Document, Stockholm: Karolinska Institute.

Penuel WR, Means B (2004). Implementation variation and fidelity in an inquiry science program: analysis of GLOBE data reporting patterns. J Res Sci Teach 41, 294–315.

Plass JL, Milne C, Homer BD, Schwartz RN, Hayward EO, Jordan T, Verkuilen J, Ng F, Wang Y, Barrientos J (2012). Investigating the effectiveness of computer simulations for chemistry learning. J Res Sci Teach 49, 394–419.

President's Council of Advisors on Science and Technology (2012). Engage to Excel: Producing One Million Additional College Graduates with Degrees In Science, Technology, Engineering, and Mathematics Washington, DC: U.S. Government Office of Science and Technology.

Ruiz-Primo MA (2006). A multi-Method and Multi-Source Approach for Studying Fidelity of Implementation, Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Seidel SB, Tanner KD (2013). "What if students revolt?"—considering student resistance: origins, options, and opportunities for investigation. CBE Life Sci Educ 12, 586–595.

Southerland SA, Nadelson L, Sowell S, Saka Y, Kahveci M, Granger EM (2012). Measuring one aspect of teachers' affective states: development of the science teachers' pedagogical discontentment scale. School Sci Math 112, 483–494.

Sunal DW, Hodges J, Sunal CS, Whitaker KW, Freeman LM, Edwards L, Johnston RA, Odell M (2001). Teaching science in higher education: faculty professional development and barriers to change. School Sci Math 101, 246–257.

Tanner KD (2011). Reconsidering "what works." CBE Life Sci Educ 10, 329–333.

Theobald R, Freeman S (2014). Is it the intervention or the students? Using linear regression to control for student characteristics in undergraduate STEM education research. CBE Life Sci Educ 13, 41–48.

Trigwell K, Prosser M, Ginns P (2005). Phenomenographic pedagogy and a revised "approaches to teaching inventory." High Educ Res Dev 24, 349–360.

Turpen C, Dancy M, Henderson C (2016). Perceived affordances and constraints regarding instructors' use of peer instruction: implications for promoting instructional change. Phys Rev Spec Top Phys Educ Res 12, 010116.

Turpen C, Dancy M, Henderson C, Singh C, Sabella M, Rebello S (2010). Faculty perspectives on using peer instruction: a national study. Paper presented at the Physics Education Research, held 21–22 July 2010, in Portland, OR.

Turpen C, Finkelstein ND (2009). Not all interactive engagement is the same: variations in physics professors' implementation of peer instruction. Phys Rev Phys Spec Top Educ Res 5, 020101.

Vickrey T, Rosploch K, Rahmanian R, Pilarz M, Stains M (2015). Research-based implementation of peer instruction: a literature review. CBE Life Sci Educ 14, es3.

Walter E, Beach A, Henderson C, Williams C (2014). Describing instructional practice and climate: two new instruments. Paper presented at the Transforming Institutions: 21st Century Undergraduate STEM Education Conference, held 24 October 2014, in Indianapolis, IN.

Walter EM, Henderson CR, Beach AL, Williams CT (2016). Introducing the Postsecondary Instructional Practices Survey (PIPS): a concise, interdisciplinary, and easy-to-score survey. CBE Life Sci Educ 15, ar53.

Wieman C, Gilbert S (2014). The Teaching Practices Inventory: A new tool for characterizing college and university teaching in mathematics and science. CBE Life Sci Educ 13, 552–569.