Design and Use of a Systematic Site Visit Protocol: Implications for Novice Evaluators and Mentors

Joy R. Lile
Jeffrey M. Flesch
Mary E. Arnold
Oregon State University

Site visits are frequently used by evaluators to gain first-hand experience and knowledge about program implementation. However, few peer-reviewed articles describe the procedures used for designing and conducting site visits. This article describes the process of constructing and using a systematic site visit protocol. Theories and concepts of evaluation, including the measurement of fidelity and quality and the importance of context to site-level implementation, guided the construction of this protocol. Using a systematic method for program inquiry can improve the consistency of qualitative observations of program activities by enhancing intentionality, transparency, and emergence within the site visit process. The method presented may be especially helpful to novice evaluators and their mentors in learning about and teaching the process of conducting site visits.

Keywords: site visits, program implementation, teaching evaluation, methodology, pedagogy

Introduction

Personal visits to program sites can be an effective method for collecting information in program evaluations. By conducting site visits in a standardized manner, evaluators can improve consistency and organization of data collection and come to a deeper understanding of the factors influencing program implementation across sites (Lawrenz, Keiser, & Lavoie, 2003).

In 2014, an evaluation was conducted of a pilot program taking place in five states. The program provided school-aged children with nutrition education consisting of several required components. The purpose of the evaluation was to identify program strengths and concerns in concordance with required program components and to develop a program model of best practices that the funder could use for taking the program to larger scale. The funder explicitly required that the evaluation team conduct site visits to programs in each pilot state.

Direct correspondence to Joy R. Lile at joy.lile@wsu.edu

The evaluation team consisted of one faculty principal investigator (PI) and two graduate student research assistants (RAs). While the PI was well-versed in the process of conducting site visits, the RAs were not, and RAs reviewed current best practices for conducting site visits. The review revealed a relative dearth of peer-reviewed literature related to designing and conducting systematic site visits. Novice evaluators and students, in particular, could benefit from new publications relating to site visit methodology. The purpose of this paper is to provide a review of current literature related to conducting program site visits, present the method used to develop a systematic site visit protocol, and share the lessons learned about conducting site visits that may be relevant to novice evaluators and their mentors.

Review of the Literature

Three important concepts can inform the development of a well-organized site visit protocol. These include local *context* at the site level, *fidelity* of the local program to the program model, and *quality* of the intervention produced. These three concepts are discussed individually and then considered as interacting elements that evaluators should take into account as they plan and implement site visits.

Context

Context is an important, though often ignored, factor in program implementation (Conner, Fitzpatrick, & Rog, 2012; Kirkhart, 2011; Rog, 2012). Variability in implementation is commonly seen in multisited programs, because site-level programs often require adaptation to local needs and resources (Lawrenz et al., 2003). Site-level institutional culture, political and social dynamics, and power differentials can create differences in implementation; these factors may be difficult for site-level implementers to detect (Kirkhart, 2011). Evaluators should analyze how the physical, organizational, social, cultural, traditional, political, and historical contexts in which a program takes place affect its implementation, and subsequently, outcomes (Conner et al., 2012). By visiting sites in person, evaluators can assess ways in which site-level program contexts interact with the delivery, and ultimately, the outcomes of the program.

Fidelity

Fidelity is important in the use of evidence-based interventions in which strict adherence to a program model is often necessary to produce anticipated outcomes, as well as in *accountability* evaluations aiming to measure adherence to external requirements like funder deliverables (Patton, 2008). Funders are often highly interested in the fidelity of a site to the program model because meeting stated requirements for delivery is often a minimum criteria for continued funding (Mowbray, Holter, Teague, & Bybee, 2003). In the past, issues of fidelity have been overlooked in evaluation; in a seminal review of 162 program evaluations published between

1980 and 1994, only 32 (19.7%) documented program fidelity, and only 13 (8%) included an analysis of how fidelity interacted with program outcomes (Dane & Schneider, 1998). Recent writers focus more on fidelity, especially when evaluating large-scale projects implemented across multiple sites (e.g., Zvoch, 2012; Zvoch, Letourneau, & Parker, 2007). The need for measuring fidelity will vary across program models, but higher fidelity is often associated with improved participant outcomes (Dane & Schneider, 1998).

In determining appropriate levels of fidelity, it is important to balance the need for strict adherence with allowances for flexibility within a program model (Barth, 2004; Dane & Schneider, 1998). In multisite programs with highly regimented models of delivery, fidelity should ideally be high and consistent across sites to produce meaningful results in the outcome evaluation (Esbensen, Matsuda, Taylor, & Peterson, 2011; Melde, Esbensen, & Tusinski, 2006). However, fidelity may be less relevant to exploratory or pilot programs that necessitate innovation.

The literature identifies five core elements of fidelity: *exposure*, *adherence*, *delivery quality*, *program differentiations*, and *participant responsiveness* (Bickman et al., 2009; Dane & Schneider, 1998). In addition, Century and colleagues (2010) point out that fidelity can also be subdivided into *structural* (program framework and organization) and *process* (relationships and interactions between key players like staff and clients) components, helping to further elucidate differences between program objectives and site-level execution. The measurement of fidelity within a program setting is dependent on the evaluation questions of the project (Patton, 2008) and should be considered from the early stages of evaluation development.

Fidelity is an important and complex factor in program implementation and a relatively straightforward concept to measure (Barth, 2004). It can be measured objectively using instruments designed in advance or by directing closed-ended questions at site-level program providers or participants. Measures of fidelity may include checklists of program requirements, participation rates and participant demographics, and surveys or interviews with participants or implementers (Bickman et al., 2009; Mowbray et al., 2003). In addition, the use of scales and quantitative measures of fidelity is growing increasingly common and allows for analyses of fidelity as a site-level variable in multilevel modeling (Resnicow et al., 1998; Zvoch, 2012; Zvoch et al., 2007), although determining the validity of these scales may be a challenging process (Mowbray, Bybee, Holter, & Lewandowski, 2006). Evaluators must work with funders and implementers to establish criteria for program fidelity, including the indicators to be used, the data to be collected, and how the indicators will be assessed for reliability and validity (Bickman et al., 2009; Lawrenz et al., 2003).

Quality

Program *quality* is a more subjective concept and necessitates measurement by an outside, objective observer (Brandon, Taum, Young, Pottenger, & Speitel, 2008). Quality describes how program presentation serves to enhance or diminish outcomes. Assessing quality is arguably more challenging than assessing fidelity because doing so requires understanding the "degree of excellence" with which a program is implemented or the relational dynamics between providers and recipients of the program (Barth, 2004). To assure objectivity, quality should be measured by external evaluators through direct observation, through interviews with staff and participants, and/or through other appropriate mechanisms such as document analysis (Brandon et al., 2008). Because quality is a difficult construct to measure, evaluators must carefully identify and justify their criteria for passing judgment on programmatic quality (Brandon et al., 2008).

Measuring Context-Fidelity-Quality Interactions

The *context*, *fidelity*, and *quality* of a program interact to affect program outcomes because the context of a program can constrain or enhance its levels of fidelity and quality. Program quality should be compared with program fidelity to understand how both interact to impact program outcomes (Barth, 2004; Dane & Schneider, 1998). A site-level program can simultaneously have a high level of fidelity and a low level of quality, or vice-versa, depending on interactions between the program model and the context of the site. For example, in evaluating a national school-based intervention Zvoch and colleagues (Zvoch, 2012; Zvoch et al., 2007) found that high implementation fidelity at the site level was correlated with average student outcomes, while both high and low student outcomes were correlated with low implementation fidelity. In this case, the program as designed may have had a low or moderate level of quality, so those sites with high fidelity also had lower quality, and some sites with low fidelity were able to improve outcomes specifically because they diverged from the program model. It is important to assess what differences exist across implementation sites; high-quality/low-fidelity programs may be a sign that implementation guidelines and expectations are not clear, while high-fidelity/low-quality programs may be a sign that the program model should be redesigned or training should be provided to enhance quality. Recording data about contextual factors influencing program quality and fidelity can help evaluators explain implementation and outcome differences across local program sites.

A foundational book edited by Wholey, Hatry, and Newcomer (2004) provides a framework for organizing a systematic method for collecting site visit data. The authors note that the objectives of site visits should focus on describing and explaining the situations of the sites within the context of the program, thereby focusing on interactions between context, fidelity, and quality. The authors suggest a careful choice of a framework in order to provide structure and avoid overcollecting data, though the framework chosen is highly dependent on the evaluation

questions. Unfortunately, few evaluators have described their own process for conducting standardized site visits in peer-reviewed journals, so few examples exist to guide nascent evaluators in planning and conducting site visits.

One example of a description of the creation and utilization of a systemized protocol for site visits can be found in Paddock and Dollahite's (2012) article. The authors describe the process of creating a highly standardized and systematic strategy to measure fidelity and quality in a large federal nutrition education quality assurance program. The protocol included evaluation questions, preferred respondents, open- and closed-ended interview questions, and a quantitative tool for checking off elements of the program model known to contribute to quality. Examples such as this can help students and early-career evaluators understand how systematic protocols are developed and used in real-world settings.

This paper describes the process of designing and implementing the site visit component of a multisite evaluation. Because the evaluation was carried out on a rapid timeline, creating and following a systemized site visit protocol enhanced the process of implementation evaluation and the evaluators' understanding of site-level programs' *contexts, fidelity,* and *quality.* Furthermore, the process of designing and using the protocol served to provide educational training for novice evaluators through a structured and hands-on evaluation experience (Tourmen, 2009). The protocol detailed here is not meant as a prescriptive tool but rather a process that evaluators can engage to develop individualized systematic protocols for a wide variety of evaluation studies.

Methods

The systematic site visit protocol process was utilized in evaluating a nutrition education program designed for a national youth-development organization and piloted in 5 states in 2014. The program engaged teens and adult educators to teach youth ages 8 to 12 about nutrition and physical activity in a 10-hour series of classes. Incorporating teens as partners in teaching was an important and novel component of this program. Grantee states were allowed to select curricula for nutrition education, as well as for training teen teachers. Each state was expected to design and provide the program to 2,500 youth during nine months. A request for proposals to evaluate the program included a quantitative outcome evaluation of all 12,500 youth participants and a qualitative implementation and outcome evaluation with visits to implementation sites.

The full project timeline was fairly compressed (March – November) for a large-scale program. Sites submitted interim reports to the funder on June 1st, after which the evaluation team had through September 30th to design, carry out, and report on the site visit process. The protocol was designed to provide evaluators a bird's-eye view of the program in each state by meeting with select administrators and personnel and observing a small number of lessons in action.

Development of the Protocol

The evaluation team was comprised of one experienced evaluator (the PI) and two graduate student research assistants (RAs). RAs began to develop the site visit protocol in early June, and a draft site visit protocol was presented to the funder on June 30th. The funder's request for applications (RFA) formed a guiding framework for the evaluation by providing an extensive and specific list of program deliverables from which the evaluation team worked. The RAs searched the literature for information on site visits, as well as on best practices in teen teaching and nutrition education programs to more effectively measure program quality.

The next step was to develop logic models for intended programmatic impacts. An overarching program-level logic model was created to describe the process by which the program was intended to change nutrition and physical activity behaviors through nutrition education. A second model was created to detail intended developmental outcomes for teens involved in teaching younger youth. State-level logic models were also created based on information provided in both state funding proposals and state interim reports, to thoroughly describe the activities being carried out or intended to be carried out in each state. (Logic models are not included because this paper focuses on methods rather than findings of the evaluation).

The team next developed a list of general evaluation questions covering conceptual aspects of implementation including the program's *adherence, delivery*, *dose, participant responsiveness*, state-level *differentiation,* and state-level program *quality* (Bickman et al., 2009; Dane & Schneider, 1998); for example: "Has the program adhered to the goals and outcomes set forth in the program plan (theory); and according to the [funder] deliverables?" The team then used the RFA to create more specific evaluation questions aligning with funder deliverables and based on fidelity/quality constructs. Below is an example of one stated goal of the funder, and the four specific evaluation questions developed based on that goal:

> *Goal 1*: Impact 12,500 underserved/at-risk youth and their families in five states with quality nutrition, budgeting, and food skills education.
>
> > *Question 1:* What recruitment strategies are most effective at reaching underserved families?
> >
> > *Question 2:* What partnerships were leveraged to recruit underserved families?
> >
> > *Question 3:* What were the most and least effective program delivery strategies to enhance responsiveness and reach the outcomes of quality nutrition, budgeting, and food skills education in youth aged 8 to 12?

Next, evaluators analyzed the evaluation questions and created a list of "preferred respondents" at each site in order to increase standardization across site visits. In each state, intended respondents included a state-level administrator, a site-level administrator, a curriculum specialist, a site-level adult teacher, a site-level teen teacher, and a youth participant and his or her parents or family if possible. Intended activities also included observing a lesson in action and reviewing the site's curriculum. Evaluators provided the list of intended activities and respondents to states before visiting, and state administrators identified individuals who best matched those roles and scheduled time with each type of preferred respondent. The process carried out for identifying respondents is similar to that detailed by Wholey et al. (2004).

Interview questions for each respondent were developed by comparing preferred respondents/ activities to evaluation questions; 36 questions were included. Data collection tools were created to take notes and organize conversations with preferred respondents. Evaluators used the list of interview questions to create note-taking pages divided into quadrants to organize notes. For example, the "curriculum" page was divided into four quadrants labeled "successes," "challenges," "modifications," and "fidelity." Figure 1 provides an example of a note-taking page.

Figure 1. Example Note-Taking Page

Program Administration	
What parts of the program model (as specified in the RFA) have worked well?	**What parts of the program model (as specified in the RFA) have been challenging?**
Data Collection:	**Programmatic Partnerships:**

The interview guide for conversations with teen teachers was more regimented to assure that the evaluators discussed the program with teens in a developmentally appropriate manner, considering that teens were also an intended recipient, as well as a provider, of the program. The teen teacher interview included eight specific questions such as, "What was the greatest thing about being a teen teacher?" and "What was tough about being a teen teacher?"

The evaluation team also developed a checklist of programmatic deliverables listed in the RFA on which evaluators could check "yes," "no," or "on-track" for each deliverable in each state, depending on whether the state-level program had already met or was prepared or unprepared to meet the deliverable in the allotted time frame. For example, during a visit in July, a program could be "on track" to meet the deliverable of reaching 2,500 students by the end of August. See Figure 2 for an example of part of the deliverables checklist.

Figure 2. Example Deliverables Checklist

Required Components	Yes	No	On-Track	Comments
Program Recipients				
Reach 2,500 underserved youth and their families				
Programs in rural, suburban, urban areas				
Program Components				
Actively engages youth and families				
10 hours of programming delivered				
2 capstones delivered				
5 hours of community service recommended				
Common Measures evaluation utilized				

The complete data collection instruments consisted of six pages, with sections on program administration, curriculum, teens as teachers, family engagement, the interview guide for teen teachers, and the deliverables checklist. Along with the data collection instruments, the site visit protocol also included information intended to help the evaluator conduct the site visit efficiently and effectively. The sections of the complete site visit protocol included grant deliverables and research questions, preferred respondents, respondent questions, state-level logic models, and data collection instruments. The state-level logic models were accompanied by a table describing administrators' names, titles, and position descriptions, and attached were the site's original application and interim reports for review. This was intended to provide the evaluator a chance to prepare for the individual site they were visiting and plan what specific questions and probes they might have for site respondents.

The site visit protocol was submitted to the funder on June 30th, and the first site visit was conducted July 6th. The protocol was revised for ease of use after feedback from the first site visit, and subsequent site visits were conducted July 22nd, July 23rd, August 6th, and August 14th. The draft evaluation report was due on September 30th. In all, during this short time, site visits were conducted in five states at nine distinct program sites.

Modifying the Protocol

The protocol allowed the evaluators to approach site visits in a systematic and organized way; however, the realities of field work necessitated changes to the protocol to enhance its utility.

After piloting the protocol at the first site, the most noticeable issue was in the note-taking pages. This instrument was thorough but long and unwieldy. The original version was organized by the individuals and roles that evaluators planned to interview at each site. However, multiple individuals often had overlapping responsibilities and areas of knowledge. As such, the note-taking pages were redesigned to cover topical areas, rather than focusing on preferred respondents. This way, the evaluator could cover one topic at a time, moving through the whole instrument with each respondent or moving between topics as the respondent changed subjects. The instrument was used to organize notes but was not followed strictly during visits, as differing circumstances at each site required rapid adaptation of data collection and organization strategies.

Recording Data

Visits were conducted across one or more days, and activities included driving to multiple sites; observing lessons in action; having lunch, dinner, and meetings with program administrators; viewing office and storage spaces to assess the logistical needs of the program; viewing curriculum materials and site records; and meeting with teen and adult teachers. Evaluators used time in transit before visits to review the site visit protocols, focusing on the logic models and highlighting specific questions for the site. Field notes were recorded in the designated areas of the data collection instruments, as well as in margins, backs of papers, and on separate lined paper, reflecting evaluators' tightly-packed agendas during site visits.

Evaluators created site visit reports both individually and collectively, often directly following the visit. The information provided in the protocol and the notes that evaluators took on data collection instruments helped to structure site visit reports. Reports were constructed both chronologically and thematically, beginning with a description of the site and the visit, and then focusing on thematic areas aligning with topical areas of inquiry (administration, logistics, curriculum, teens as teachers, etc.). Site visit reports included footnotes and highlights of information of particular interest or importance and questions for site follow-up via phone or email. The checklist of deliverables was completed after the site visit as evaluators reviewed site visit reports and notes. After each report was compiled, the RAs analyzed it and extracted information aligning with each evaluation question. This helped to organize information systematically and ensure that evaluation questions were answered thoroughly.

Analysis

Themes for the final report were developed both prescriptively and emergently. Emergent data analysis techniques stem from *grounded theory* (Corbin & Strauss, 2008) in which qualitative researchers allow themes to emerge organically from the experiences of their research subjects, while the use of sensitizing themes is akin to *focused coding* (Emerson, Fretz, & Shaw, 2011) in

which researchers develop a list of themes based on a predetermined theoretical and conceptual framework and apply them to the data. Sensitizing and predetermined themes were developed based on the programmatic requirements and best practices of nutrition education and teen teaching programs, as well as the required deliverables stated in the RFA. Emergent themes were compiled as site visits proceeded and served to influence the data collection process iteratively and inform final findings. As themes emerged, evaluators probed on particular questions and posed follow-up questions to previously visited sites.

The evaluation team met to create a list of primary themes for the final report and divided the list amongst team members to create sections of the final report. Sections were organized by topical themes, including teen teaching, teen training methods and curricula, nutrition curricula, programmatic partnerships, staff training, budget and staffing, and components of the grant that were under-realized across sites. For each section, team members read back through site visit reports to summarize and compare across sites and wrote on successes and challenges evident in program implementation.

The preliminary evaluation report was provided to the funder, as well as to state-level program administrators. The preliminary report was discussed at a debriefing meeting hosted by the funder and attended by the PI and program administrators from each state. During this debriefing, administrators were asked to comment on the preliminary report and provide their own perspectives of the program for comparison in order to corroborate evaluators' findings and provide further insight into the utility of the program model. Themes from the preliminary reports were refined and expanded upon during this discussion. The final report was presented to funder after quantitative outcome evaluation data was analyzed and added to the report.

Discussion

Through the process described above, the evaluation team created a systematic site visit protocol. The discussion focuses on how the team used the protocol during site visits and what lessons can be garnered from the process.

Utility of the Protocol

The components of the site visit protocol enhanced its utility on the ground, providing structure and organization to the process. Evaluators continually reviewed research questions, interview questions, and site-specific information during visits to ensure consistency of data collection and recorded notes and ideas on documents to spur further discussion.

Including personnel lists and state-specific logic models in the site visit materials increased the utility of the protocol. Logic models for each site-level program allowed for a quick review of

the program immediately preceding each site visit so evaluators could make note of particular questions for site-level administrators. Notes were made on logic models to clarify which parts of the grantee's proposal were being implemented. Notes were also made on staff position tables in order to remind evaluators of which individuals were answering questions (e.g., Nancy X, Program Assistant). Having a summary of the site's unique program model and personnel plan on hand helped evaluators prepare for and efficiently carry out each visit.

While meeting with preferred respondents, evaluators referred to the list of interview questions and also to the note-taking pages and deliverables checklist to ensure that necessary topical areas were covered. The inclusion of the complete list of interview questions in site visit protocols allowed evaluators to check off completed questions and make notes of which questions required further inquiry. The interview guide for teen teachers helped evaluators maintain a level of consistency across groups of teen respondents, and topically-organized note-taking pages allowed flexibility in the data sources while still providing a systematic organization strategy for data collection. Evaluators were able to add questions as they arose and organize new questions thematically and by respondent to ensure that they would be asked of the correct individuals. For example, one protocol includes the note, "budget and staffing" in the questions under the administration category. This topic arose as a salient issue during an early site visit and became a primary theme in the final report as different approaches were observed across sites. Having access to a complete and organized list of evaluation and interview questions during each site visit allowed evaluators to record and organize their thoughts more systematically and ensured that salient topics were discussed.

As aforementioned, the checklist of grant deliverables was used post-visit while creating site visit reports. The checklist ensured that evaluators collected all of the necessary information and allowed them to track where low fidelity existed within state-level programs. For example, one grant deliverable (providing bags of groceries to participating families) was added to the list of program requirements late. Review of the checklists highlighted that some sites had more complex operations because they were trying to reach this grant deliverable. Sites with fewer logistical challenges related to this late expectation had neglected to provide this additional deliverable or provided it in a way that did not align perfectly with the RFA requirements to reduce logistical challenges. In this way, the checklist enabled evaluators to compare deliverables envisioned by the funder with realities of time and staffing concerns for the sites to assess how fidelity interacted with quality and sustainability of the program. This finding points to the importance not only of measuring the processes of a program but also of understanding why certain processes are implemented and what motivations and contextual factors drive differences in program delivery.

Despite its careful and deliberate design, the site visit protocol was not always followed with perfect fidelity due to the realities of field work. For instance, although the protocol was well

organized, the reality of data collection was rarely a well-organized and standardized process. Evaluators met with widely varying groups of people across sites depending on the availability of participants and administrators. At some sites, teens were interviewed individually, while at others, they answered questions in groups, and at one site, only one teen was available at the time of the interview. Evaluators took notes on the interview guide, as well as on lined paper and on the backs of other sheets within the protocol, which sometimes led to disorganized notes. Widely differing program implementation and availability of individuals across sites made data collection challenging, and although a systematic plan was in place, evaluators were not always able to follow that plan.

Despite limitations, the evaluation process benefited from the inclusion of a structured protocol for conducting site visits. Although the protocol was not followed flawlessly, it provided clear guidance to evaluators as they conducted the site visits and a framework for preparing complete and organized site visit reports. The preparation involved in creating the site visit protocol allowed evaluators a deep understanding of the intentions of the program model, and having the full protocol on hand during the site visit meant evaluators could continually refer back to the program model and intended questions during the visit. Use of the protocol enhanced the objectivity and consistency with which the site visit was conducted, as well as evaluators' ability to think quickly and make analytical deductions on the ground. Each site an evaluator visits is inevitably unique from the last, so some level of variability in implementation across sites is expected. However, coming prepared for variation with a strategic and systematic plan for site visits will help evaluators collect more complete data.

Measuring Quality and Fidelity within Context

The protocol developed for this evaluation specifically addressed issues of both *fidelity* and *quality* (Bickman et al., 2009; Dane & Schneider, 1998) within each unique site *context*. Evaluation areas of inquiry tied directly to issues of fidelity, covering program *adherence;* program *delivery;* program *dose; participant responsiveness/engagement;* and *differentiation* between national-, state-, and site-level programs. The incorporation of grant deliverables in the logic model and subsequent comparison between overarching and state-level logic models enabled an understanding of how the programs both *adhered to* and *differed from* the program model presented by the funder. The data collection instruments were crafted to measure specific aspects that evaluators understood to be most important to programmatic fidelity. Visiting sites allowed evaluators to assess the extent to which *delivery* of each site-level program followed the goal of engaging teens as true partners in teaching. Evaluators were able to assess participant *dose* by probing on lesson time provided and participation rates and participant *responsiveness* and engagement through observation; quantitative outcome data also helped assess responsiveness.

The evaluators determined program *quality* based on a literature review and the research team's prior experience in designing and evaluating youth programs. A prior understanding of the literature was crucial in evaluating the quality of each site-level program. Aspects of quality reflected in the literature review included youth-adult interactions, approaches to teen teaching, and uses of nutrition education curricula. The RFA provided more empirically informed guidelines for nutrition education but fewer for teen teaching programs, so reviewing the literature on teens-as-teachers was a crucial step in measuring program quality. A thorough investigation of the literature led to the development of an overarching logic model for teen teacher programs. Quality was then measured by comparing program implementation to the necessary aspects of teen teaching according to prior research. For example, the inclusion of time for reward and recognition, as well as for debriefing and feedback are cited as important features of teen teaching programs (Lee & Murdock, 2001). Evaluators used these and other indicators from the literature to rate and describe the quality of the different approaches to teen teaching. Wide variation was found in the quality of teen teaching aspects of the site-level programs, with teens acting as lead teachers or collaborators in some programs and as assistants with few responsibilities at others. As demonstrated in previous literature (Zvoch, 2012; Zvoch et al., 2007), an overly-flexible program model for the teen teaching elements led to high differentiation in program quality and to programs with high *fidelity* but low *quality*. Having conducted a thorough literature review allowed evaluators to make specific recommendations to the sites and funder to improve program design and implementation.

Implications for Novice Evaluators

Several lessons can be taken from the process of developing a systematic site visit protocol. Such a protocol can be useful in structuring and guiding novice evaluators through the process of conducting an implementation evaluation and learning about a program.

Systemized programmatic inquiry. Utilization of a systematic site visit protocol denoting areas of inquiry, preferred respondents, and the unique program model for the site helps the evaluator ensure consistency and completeness of data by enhancing the *intentionality* with which the site visit is conducted, the *transparency* of the evaluation process, and the *emergence* of important themes and findings for evaluation reports.

Systematic protocols can enhance *intentionality* within evaluations by allowing evaluators to create and use systematic data collection instruments. Evaluators should move through this process in a logical progression: 1) Gather and review information on intended program deliverables, 2) Conduct a literature review of programmatic best practices to measure elements of program *quality*, 3) Develop program logic models, 4) Develop evaluation questions, 5) Create lists of preferred respondents, 6) Develop interview questions aligning with preferred respondents to answer evaluation questions, and 7) Draft data collection instruments. The order

of this procedure roughly mirrors the process detailed in the Methods section. By compiling these tools and resources into one document, evaluators can work to ensure they have been thorough and intentional in exploring necessary areas of inquiry during visits. This also enables evaluators to be prepared for and keep track of complexities of the program at the site level.

Systematic site visits have the potential to increase *transparency* in the site visit process. The thought of external program evaluation can be daunting for implementers. By preparing site-level administrators with a list of preferred respondents and themes for the site visit, evaluators can reduce stress on site administrators and shift the power dynamics of the evaluation process. Administrators can fully prepare for site visits by scheduling the evaluators' time and preparing documentation to answer evaluators' questions. During the aforementioned project, evaluators provided the list of preferred respondents and activities to sites in preparation for site visits. Many administrators used these items to structure evaluators' time down to the hour which was highly efficient. Providing areas of inquiry and preferred respondents to site-level administrators in advance can allow them to feel prepared for and comfortable with the site visit process.

The structure created by a systematic site visit protocol can also enhance *emergence* in the data analysis process. By maintaining detailed and organized notes on the program, evaluators can record questions and themes as they emerge and easily refer back to them during site visits. This is necessary in any qualitative research but especially important for evaluators, who often must become experts on the intricacies of complex programs within short periods of time. Becoming familiarized with the program and having a clearly organized tool for recording and revisiting information both during and after site visits helps the evaluator recognize and remember important themes as they emerge across site visits.

Site visits as a teaching technique. The creation and use of a systematic site visit protocol as described has particular relevance to the education of novice evaluators. Tourmen (2009) notes that new evaluators tend to focus on utilizing systematic and technical methods in their work, while experienced evaluators focus on programmatic goals and political implications involved in evaluation and base their evaluation approach more on usability than technical exactitude. The evaluation team for this project was comprised of one highly experienced evaluator and two students, and the experiences of these team members mirrored Tourmen's continuum.

Creating a systematic site visit protocol allowed the RAs to develop their knowledge and experience base in the field of program evaluation. During site visits, the materials in the site visit protocol helped the RAs keep track of a large amount of information and increased their level of comfort in requesting to observe the necessary components of the program. Experienced evaluators may feel comfortable going to site visits with less structured plans for data collection, but for novice evaluators, the process of preparing for field work in a systematic way is invaluable.

The processes of analysis for the RAs and the PI also differed in ways aligning with Tourmen's (2009) assessment of the different processes used by experienced vs. novice evaluators. The RAs applied evaluation and interview questions to the site visit reports, drawing out themes. The PI did this as well but also used her knowledge of youth development programs and of the special challenges site-level administrators might have in running such a program to shape the areas of inquiry and the focus of the findings. For example, the PI identified budget concerns as a primary theme during a site visit, noting how budgets were allocated differently across states and how budget allocation affected staff time, and therefore, project organization and sustainability. Having both experienced and novice evaluators on the team provided multiple perspectives from which to view the program and resulted in an analysis that was both structured and responsive to the unique situations of the state-level programs.

Conclusion

This paper presented the description of a systematic site visit protocol in an attempt to encourage the creation and utilization of well-planned frameworks for conducting site visits. Within a short time period, evaluators were able to develop a protocol including *fidelity* and *quality* constructs that enabled the collection of extensive data during brief site visits. This method is useful in conducting implementation evaluations because it promotes a logical and comprehensive procedure through which to assess implementation across multiple sites. By developing and implementing systematic site visit protocols, novice evaluators can learn to conduct organized site visits and collect credible data. By sharing and critiquing techniques and methods for conducting site visits, evaluators can promote high-quality data collection in evaluation, as well as provide effective training for students and novice evaluators.

References

Barth, M. C. (2004). A low-cost, post hoc method to rate overall site quality in a multi-site demonstration. *American Journal of Evaluation, 25*(1), 79–97. doi:10.1177/109821400402500106

Bickman, L., Riemer, M., Brown, J. L., Jones, S. M., Flay, B. R., Li, K.-K., . . . Massetti, G. (2009). Approaches to measuring implementation fidelity in school-based program evaluations. *Journal of Research in Character Education, 7*(2), 75–101.

Brandon, P. R., Taum, A. K. H., Young, D. B., Pottenger, F. M., III, & Speitel, T. W. (2008). The complexity of measuring the quality of program implementation with observations: The case of middle school inquiry-based science. *American Journal of Evaluation, 29*(3), 235–250. doi:10.1177/1098214008319175

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. doi:10.1177/1098214010366173

Conner, R. F., Fitzpatrick, J. L., & Rog, D. J. (2012). A first step forward: Context assessment. *New Directions for Evaluation*, *2012*(135), 89–105. doi:10.1002/ev.20029

Corbin, J. M., & Strauss, A. L. (2008). *Basics of qualitative research : techniques and procedures for developing grounded theory* (3rd ed.). Los Angeles, CA: Sage.

Dane, A., & Schneider, B. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psycology Review*, *18*(1), 23–45. doi:10.1016/S0272-7358(97)00043-3

Emerson, R. M., Fretz, R. I., & Shaw, L. L. (2011). *Writing ethnographic fieldnotes* (2nd ed). Chicago, IL: University of Chicago Press.

Esbensen, F. A., Matsuda, K. N., Taylor, T. J., & Peterson, D. (2011). Multimethod strategy for assessing program fidelity: The national evaluation of the revised G.R.E.A.T. program. *Evaluation Review*, *35*(1), 14–39. doi:10.1177/0193841X10388126

Kirkhart, K. E. (2011). Culture and influence in multisite evaluation. *New Directions for Evaluation*, *2011*(129), 73–85. doi:10.1002/ev.356

Lawrenz, F., Keiser, N., & Lavoie, B. (2003). Evaluative site visits: A methodological review. *American Journal of Evaluation*, *24*(3), 341–352. doi:10.1177/109821400302400304

Lee, F. C., & Murdock, S. (2001). Teenagers as teachers programs: Ten essential elements. *Journal of Extension*, 39(1), Article 1RIB1. Retrieved from http://www.joe.org/joe/2001february/rb1.php

Melde, C., Esbensen, F. A., & Tusinski, K. (2006). Addressing program fidelity using onsite observations and program provider descriptions of program delivery. *Evaluation Review*, *30*(6), 714–740. doi:10.1177/0193841X06293412

Mowbray, C. T., Bybee, D., Holter, M., & Lewandowski, L. (2006). Validation of a fidelity rating instrument for consumer-operated services. *American Journal of Evaluation*, *27*(1), 9–27. doi:10.1177/1098214005284971

Mowbray, C. T., Holter, M., Teague, G., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, *24*(3), 315–340. doi:10.1177/109821400302400303

Paddock, J. D., & Dollahite, J. (2012). Nutrition program quality assurance through a formalized process of on-site program review. *Journal of Nutrition Education and Behavior*, *44*(2), 183–188. doi:10.1016/j.jneb.2011.03.006

Patton, M. Q. (2008). *Utilization-focused evaluation*. Thousand Oaks, CA: Sage.

Resnicow, K., Davis, M., Smith, M., Lazarus-Yaroch, A., Baranowski, T., Baranowski, J., . . . Wang, D. T. (1998). How best to measure implementation of school health curricula: A comparison of three measures. *Health Education Research*, *13*(2), 239–250. doi:10.1093/her/13.2.239

Rog, D. J. (2012). When background becomes foreground: Toward context-sensitive evaluation practice. *New Directions for Evaluation*, *2012*(135), 25–40. doi:10.1002/ev.20025

Tourmen, C. (2009). Evaluators' decision making: The relationship between theory, practice, and experience. *American Journal of Evaluation*, *30*(1), 7–30. doi:10.1177/1098214008327602

Wholey, J. S., Hatry, H. P., & Newcomer, K. E. (Eds.). (2004). *Handbook of practical program evaluation* (2nd ed.). San Francisco, CA: Jossey-Bass.

Zvoch, K. (2012). How does fidelity of implementation matter? Using multilevel models to detect relationships between participant outcomes and the delivery and receipt of treatment. *American Journal of Evaluation*, *33*(4), 547–565. doi:10.1177/1098214012452715

Zvoch, K., Letourneau, L. E., & Parker, R. P. (2007). A multilevel multisite outcomes-by-implementation evaluation of an early childhood literacy model. *American Journal of Evaluation*, *28*(2), 132–150. doi:10.1177/1098214007301138

Joy Lile is a PhD candidate in Human Development and Family Studies at Oregon State University, as well as a Regional Extension 4-H Specialist at Washington State University.

Jeffery Flesch, MS, is an instructor at Oregon State University in the Human Development and Family Studies program where he received his Master's degree in 2015.

Mary Arnold, PhD, is a Professor in the School of Social and Behavioral Health Sciences and an Extension Specialist in youth development at Oregon State University where she works with the 4-H Youth Development Program. Her program planning and evaluation capacity building work with 4-H in Oregon and nationally has contributed to a change in how 4-H programs are articulated, planned, and evaluated.