Evaluation Technical Assistance Update
for OAH & ACYF Teenage Pregnancy Prevention Grantees

December 2011 • Update 3

Frequently Asked Questions: Reporting Implementation Findings

As part of the technical assistance (TA) to TPP and PREIS grantees, the Evaluation TA team will produce a series of Evaluation Updates that discuss topics relevant to the rigorous impact evaluations. Grantees' requests for TA and conversations with TA liaisons determine the topics and questions for these updates. This update features answers to frequently asked questions about the role an implementation evaluation plays in supplementing impact evaluation findings in final study reports.

IN FOCUS: High-Quality Implementation Evaluations are an Extension of OAH and ACYF Fidelity Monitoring Expectations

In July 2011, OAH and ACYF provided guidance and training on fidelity monitoring. Fidelity monitoring increases the likelihood that the funded program is implemented as intended. It is also an important component of the TPP and PREIS performance measurement system, since data collected as part of the fidelity monitoring plans will allow OAH and ACYF to describe elements of program delivery and receipt consistently across all funded grantees.

This update addresses a complementary topic—high-quality implementation evaluations designed to supplement the rigorous, independent effectiveness evaluations being conducted across Tier 1 C/D, Tier 2, and PREIS grantees. The descriptive findings emerging from high-quality implementation evaluations can help readers understand the impact findings and generate hypotheses about why the program did or did not have a positive impact. A good implementation evaluation report will do this in three ways. First, it will describe the program that was implemented and the degree to which it adhered to the intended program using data from the fidelity monitoring. Second, it will describe the extent to which the experiences of the treatment and control groups differed. Third, the findings will help readers understand more about the context in which the program was implemented and evaluated.

A scientifically conducted implementation evaluation is one that produces reliable and unbiased results. Sound, scientific descriptive evidence is essential for generating valid hypotheses about why a program is or is not effective. To acquire such evidence from an implementation evaluation, the independent evaluator should have primary responsibility for designing and conducting it, including developing new instruments and collecting data. However, this may not always be possible for all aspects of the implementation evaluation. For example, in order to collect the OAH and ACYF required number of observations within budget, TPP and PREIS grantees may need to rely on grantee staff. In addition, there may not be sufficient resources to collect any additional implementation data beyond the amount required for the performance measures.

Study reports should, therefore, be very clear about these limitations of implementation evaluation findings. For example, if none of the observations were conducted by an independent evaluator, the report findings cannot be considered free from observer bias and the report should indicate this. And if the evaluator cannot collect information on the services received by the control group, or additional services received by the treatment group, then the contrast between the two groups cannot be fully described in the report. In short, the report should clearly document what is known—and more importantly, what is unknown—about program adherence, contrast, and context. Evaluators should not ignore limitations of the evidence when drawing conclusions, and should not speculate in the absence of reliable and unbiased information.

Office of Adolescent Health

Family and Youth Services Bureau

What Does An Assessment of Program Adherence Include?

Describing adherence to the program involves documenting: What was offered and what was received? Who attended the program? Who provided the program and how was it delivered? These questions can be answered with a description of the dosage, the content, and the means by which the program was provided.

Dosage incorporates two dimensions: what is offered and what is received (attendance). For most interventions, describing what is offered will involve reporting the number of sessions delivered, their average duration, and their average frequency (such as once a week or once a month). Evaluators can use youth attendance rates to describe dosage in terms of what the program youth received and who actually attended the program. The report should also clearly state what percentage of the program group did not receive any services (no-shows).

The content of the program delivered to youth is another aspect, along with the data on dosage, of what was offered to program youth. To understand what content is provided, it is important to describe the degree to which the planned sessions and activities were delivered. For TPP and PREIS grantees, the OAH-approved facilitator logs are the most likely way to collect the data needed for this assessment.

To answer who provides the program and how it is provided, the report should describe the means by which the program is delivered. Each program's expected means of delivery is unique. For example, some programs are expected to be delivered by individuals with particular qualifications; in these cases, the desired staff qualifications and those of the staff actually hired to provide the program should be described in the report. Other programs specify the use of particular pedagogical approaches or focus on developing mentoring relationships between facilitators and participants. In such cases, the report should include an assessment of the degree to which these approaches or relationships unfolded during the delivery of the program. For programs that rely on technology, such as the delivery of messages or content via the internet or video, the report should describe any challenges using the hardware or software.

OAH and ACYF expect that fidelity monitoring will collect data on the core program components (content, pedagogy, and implementation). This means that TPP and PREIS grantees will have a lot of information that evaluators can use in the final report to describe and assess dosage, content, and means of program delivery. Grantees and their evaluators are encouraged to use the reporting framework presented above; however, they should consider whether additional data collection would be useful to provide a more complete description of program adherence. If grantees and evaluators are interested in identifying additional data that could be collected and best practices for collecting them, they should contact their TA Liaison.

OAH and ACYF also expect that the information that is collected as part of fidelity monitoring will be used to continuously improve program implementation. It is imperative for the evaluator to be aware of program implementation adjustments made based on this continuous feedback, and the data that informed the adjustments. Any changes to program implementation and the basis for making them should be thoroughly described in the final report.

Why and How Should I Describe All TPP-Related Experiences of the Treatment and Control Groups?

Describing the participation of the treatment and control groups in TPP-related experiences (apart from the program being tested) is important for understanding the degree of contrast in services and experiences between the two groups. For example, the report could describe the degree to which program and control youth received information about the outcomes of interest (such as abstinence, birth control, and sexually transmitted infections) and the source of the information (such as other TPP programs, health class, church, community center, doctor, friends, or parents). Youth survey responses are a good source for this information, including the names of other programs in the area in which youth are participating. Interviews with staff involved in the TPP program being evaluated can also provide information about the range of services within the community and how they may differ from one another and from the tested intervention.

Should I Describe Adherence for Programs Offered to the Control Group?

That depends on the counterfactual condition. If the evaluation is comparing the TPP grant-funded program to another teenage pregnancy prevention program, or to another program chosen specifically for the counterfactual condition (that is, not a range of "business as usual," but a distinct program), then the answer to this question is "yes." In this case, it is important to describe adherence for the control program just as it is for the treatment program: Information on dosage, content, and means of delivery will help define the differences between what services treatment group youth received, relative to those in the control group.

If the evaluation is comparing the program to "business as usual" in schools, clinics, or communities, then the answer to the above question is likely "no." Testing the program against “business as usual” assumes that there will be variation in the degree to which the “usual business” is implemented as expected. That variation is part of the reality of the counterfactual condition.

What Aspects of Context Should Be Described?

Multiple factors in the organizations and communities operating TPP and PREIS evaluations can influence the implementation of the program being tested and the conduct of the evaluation. Often, these factors are beyond the control of the program staff and the evaluator. Still, they are important to identify and report because they may help generate hypotheses regarding what contributed to or hindered adherence to the program model. For example, when reporting on context, the evaluator might describe the introduction of a second, different teenage pregnancy prevention program in evaluation schools, an unanticipated delay in program implementation, or a principal change at an evaluation school that may have altered the school’s receptivity to the program and/or the evaluation.

Determining which contextual factors to report will be unique to each program and evaluation. Like the rest of the implementation evaluation, this depends on how and where the program is provided, what services (if any) the control group is receiving, and the number of participating sites. Evaluators may take a limited perspective on context if the intervention occurs in one organization or community, while a broader perspective may be appropriate if multiple organizations or communities are involved. If grantees and evaluators are interested in identifying additional data that could be collected to describe context and best practices for collecting them, they should contact their TA Liaison.

Members of the Eval TA team are available to provide individualized feedback to grantees on implementation evaluation plans. If you are interested in identifying additional implementation data that could be collected and best practices for collecting them, please contact your TA liaison.

How Can I Improve the Scientific Quality of the Implementation Evaluation?

For a high-quality implementation evaluation to provide reliable and unbiased findings, the data must be gathered scientifically. Some strong recommendations for ensuring high-quality implementation data are offered below; however, these are offered with an appreciation of the fact that there are no specific HHS evidence standards for the findings from the implementation evaluation, as there are for the evidence from your impact evaluation. Also, these recommendations are made with an understanding that some TPP and PREIS grants might not have the resources to conduct all of these activities. In these cases, evaluators should note the limitations of the implementation evaluation findings and factor in those limitations when drawing conclusions.

1. ***Oversee instrument development.*** The evaluator is in the best position to finalize instruments and protocols that will support data collection for the implementation evaluation, because evaluation training includes how to collect reliable information across multiple data collectors. This does not preclude the use of instruments that have not been developed solely by the evaluator, such as the fidelity monitoring logs or developers’ observation forms. However, if additional data collection is feasible it would be good practice for the evaluator to develop any supplemental instrumentation, such as observational scales that capture specific pedagogical approaches or protocols for staff interviews.

2. ***Provide data collection guidance and assess the completeness and accuracy of the data.*** The data for the implementation evaluation will likely be collected by several different groups of people, including program staff, program facilitators, and members of the independent evaluator’s team. To ensure high-quality data are collected consistently and accurately, evaluators can oversee data collection and conduct interim assessments of data quality.

   Some of the data that the evaluator will use for the implementation evaluation, such as attendance records, facilitator logs, and records of staff qualifications, will be collected by program staff. Program

(continued)

staff may have limited experience collecting data for an evaluation, and evaluators could provide training and instructions for collecting accurate and complete data. For example, program staff might benefit from written directions and training on how and when to fill out attendance records and other forms, such as facilitator logs. Evaluators are in a better position to anticipate all the various scenarios data collectors will face, and provide written guidance for how to code data in each situation to ensure consistency within and across data collectors.

In addition, evaluators should request that data from all collectors, including program staff and members of the evaluation team, be submitted at regular intervals for the evaluator to assess, and conduct an assessment of the data. For example, the evaluator can examine the data to identify any patterns of missing data or obvious inconsistencies. Possible probes to evaluate the level of missing data include: Are particular data elements missing more frequently than others? Are data from some facilitators/sites/ data collectors more likely to have less complete data? Are the requested details being provided? Are the unexpected codes or response values being used? The evaluator can also ensure that observers achieve acceptable levels of reliability (with other observers or with an anchor) before going into the field. When observations are being conducted over a long period of time, periodic inter-rater reliability checks are recommended.

3. ***Select a representative, not purposive, sample of sessions for observation.*** It is unlikely that one would want to—or have the resources to—observe all of the implemented program sessions across all lessons, sites, and facilitators. Therefore, some sub-sampling will be necessary. This sub-sampling has two stages: identifying a sub-sample of the intended lessons, and identifying a sub-sample of the sessions through which these lessons will be delivered (across sites and across facilitators).

It is useful to engage the program developer or program staff in the selection of the intended program lessons for observation. Curriculum material can also be used as a resource for creating a sampling plan. These resources should guide the selection of observations toward the more substantive content that connects with the program's theory of change. For example, program staff may suggest that particular lessons, such as introductory or concluding lessons, are not good for observation because they do not deliver program content. Considering lesson content, therefore, means the observations can focus on more substantive lessons.

Once the program lessons for observation are identified, the evaluator can assist by guiding the selection of the program sessions that will deliver these lessons. It is important that specific sites or facilitators are not purposively selected, and that sessions are not selected based on convenience of a site's location (in multi-site interventions) or schedule (for example, weekday or Saturday). Purposive or convenience sampling may lead to an incomplete picture of program delivery across sites and facilitators. The evaluator can assist by developing a sampling plan that will produce a representative sample of sessions and facilitators within each site.

If evaluators conduct a subset of all observations (with the remainder conducted by grantee or program staff), a similar process should be used to select the sessions for that subset. This is important to avoid any systematic differences between what the evaluator and others observe. For the final report, the evaluator should consider whether to analyze only the data from the truly independent, evaluator-conducted observations or to use data from all the observations. Each approach has limitations—the evaluator may have resources to observe only a small proportion of all sessions, but observations by program or grantee staff cannot be considered free from observer bias.

This update was written by Susan Zief, Juliette Henke, and Heather Zaveri from Mathematica Policy Research and Karen Walker from Child Trends for the HHS Office of Adolescent Health under Contract #HHSP23320095642WC, Task Order No. HHSP23337017T. Input was provided by Ellen Kisker of Twin Peaks Consulting, LLC and Jean Knab and Russ Cole of Mathematica.