EVALUATION TECHNICAL ASSISTANCE BRIEF
for OAH & ACYF Teenage Pregnancy Prevention Grantees

July 2011 • Brief 1

Planning Evaluations Designed to Meet Scientific Standards: Communicating Key Components of the Plan for a Rigorous and Useful Evaluation of a Teenage Pregnancy Prevention Program

In 2010, the U.S. Department of Health and Human Services (HHS), under a contract with Mathematica Policy Research, initiated the Pregnancy Prevention Research Evidence Review (PPRER) to identify rigorously evaluated and effective teenage pregnancy prevention program models. The review effort has established scientific standards to assess the credibility of evaluation findings, and uses those standards to assess completed evaluations of programs designed to improve teenage outcomes related to sexual activity, contraceptive use, sexually transmitted infections (STIs), pregnancy, or birth. Each completed evaluation is rated as having high-, moderate-, or low-quality evidence. Low-quality evidence is not used to determine program effectiveness.

The standards that guide this review effort, available on the website for the Office of Adolescent Health (OAH) (http://www.hhs.gov/ash/oah/oah-initiatives/tpp/eb-programs-review-v2.pdf), assess the internal validity of the observed program impacts. In other words, the standards assess the degree to which the observed effects of the program are credible and whether there is a high probability that the program caused them. Key areas of assessment include the loss of the study sample over time (sample attrition), equivalence of the analytic sample, and whether any factors associated with the data collection or the delivery of the program model confound the ability to attribute impacts to the program model. In general, well-implemented randomized controlled trials (RCTs) are rated as providing high-quality evidence, whereas well-implemented quasi-experimental designs (QEDs) and RCTs with design flaws are rated as providing moderate-quality evidence.

The review effort identified 28 program models that improve youth outcomes relating to sexual activity, contraceptive use, sexually transmitted infections, pregnancies, or births. The review effort is ongoing, meaning that HHS will update the review findings periodically with evidence that meets the standards; additional effective program models could be identified. In addition to identifying evidence-based program models, the review effort also highlighted the large number of evaluations that did not meet the evidence standards and could not be used to assess program effectiveness. Of the 199 evaluations first identified as relevant to the review effort, more than half (106) received a low evidence rating.

The availability of the HHS evidence standards, and the fact that HHS will continue to use them to assess completed studies of teenage pregnancy prevention programs, provides evaluators with a framework for designing impact evaluations that could be rated as moderate or high, and program staff with an incentive to commission such evaluations. Currently, HHS expects that the grantee-level effectiveness evaluations conducted by its Teenage Pregnancy Prevention grantees will provide evidence that meets the standards.

Evidence quality alone is not the sole determinant for whether evaluation findings will be useful to decision makers. A useful evaluation is also one that answers policy-relevant questions about program effectiveness. An evaluation plan has a higher probability of resulting in a completed study that meets scientific standards and is useful if it (1) describes the program model and articulates research questions that are of interest to decision makers, (2) demonstrates that program impacts could be detected, (3) uses a study design that will provide valid estimates of program impacts, and (4) describes plans to collect and analyze data on context and implementation so that findings can be understood.

This brief discusses planning effectiveness evaluations that will meet both objectives—providing rigorous evidence that will meet HHS evidence standards that will also be useful to decision makers.

Describe the Program Model and Its Relationship to the Primary Research Questions and Outcomes

Evaluation plans should describe all components of the program to be implemented and tested, present a theory of change that links the provision of these components with the intended outcomes, and identify research questions that test hypotheses established in the program's theory of change.

Theory of Change

Documenting a program's theory of change is usually accomplished through a visual representation, such as a logic model. Understanding the theory of change is critical for assessing all aspects of the evaluation design. The theory of change should establish the fundamental components of the program model and how they interact with one another to affect intermediate and long-term outcomes.

Activities and Services

The evaluation should be timed to work in concert with program implementation plans. To assess this, the plan should provide details on specific activities or services of the program to be evaluated, such as when, where, and how they are delivered, and the intended dosage. It also is important to describe any modifications being made to an existing program or curriculum.

Program Participants

The eligibility criteria for program participation are important for understanding whether the evidence can be used to make claims about program effectiveness. Plans should document whether program participation is voluntary or mandatory for those eligible, and how program participation decisions will be used to define the study samples.

Primary Research Questions

The primary research questions should convey the key hypotheses the evaluation will test. These hypotheses should be aligned with the program components and the theory of change, including the timing of when key outcomes could be observed.

Demonstrate that Program Impacts Could Be Detectable

The evaluation plan should also demonstrate that the evaluated program is capable of having impacts on key outcomes and that the study design has a good chance of detecting them. More specifically, the plan should describe: the ability to implement the program with fidelity; the statistical power of the study to detect program impacts on key outcomes; the degree of contrast in experiences of the program (treatment) and comparison groups; and, the potential for "contamination," in which comparison group members are exposed to program services or messages.

Implementation Fidelity

Program effectiveness evaluations are implemented to determine if the program being tested indeed works. Impact estimates are therefore more useful if the program model is actually implemented as intended. Evaluation plans should include evidence that the program has been, and can continue to be, implemented with fidelity, meaning that all components and services can be implemented in a manner that is consistent with the intentions of the program developer. Plans should describe the steps that will be taken to ensure fidelity when a program is implemented in new sites.

Statistical Power

Assessing a study's power to detect impacts is not straightforward—it is part science and part judgment. Power calculations should reflect the expected size of the sample (or subsample) at the time that key outcomes will be measured, the anticipated prevalence or level of the outcome in the comparison group, and any necessary adjustments for the clustering of individuals within groups. The plan should also discuss whether the difference between the two groups that can be detected as statistically significant would be expected given the nature of the program (for example, a two-day workshop versus an intensive intervention that extends over several years) and any prior evaluation findings.

The Comparison Condition

No matter how well the program is implemented or the size of the sample, an impact might not be detected if the program group and comparison group receive similar services and activities. In the field of public health, and especially teenage pregnancy prevention, a school or community could already be saturated with programming designed to affect outcomes similar to those being examined in the planned evaluation. Documenting any other related services that are available and the extent of saturation is an important component of an evaluation plan. The extent of saturation that is acceptable depends in large part on the evaluation objectives and key research questions.

Contamination of Comparison Condition

When program and comparison group members are in close proximity, concerns arise that comparison group members might access program services or be indirectly exposed to them through interactions with the program group. Such contamination of the comparison group will diminish the ability to detect the true impact of the program model. Thus, plans to minimize and measure such contamination should be clearly described, especially when individuals (as opposed to schools or community sites) are assigned to the program and comparison groups.

Use a Study Design that will Provide Valid Estimates of Program Impacts

A key consideration when assessing plans is whether the study will provide valid, unbiased estimates of program impacts. Regardless of whether the sample design proposed is an RCT or QED, plans should clearly document how the evaluators will (1) establish the internal validity of the study by forming program and comparison groups, (2) maintain the design and minimize

bias in the study as it unfolds, and (3) avoid completely aligning a single aspect of the evaluation design (for example, one unit of program delivery or the mode of data collection) with either the program or the comparison group. The plan should also outline a logical data collection plan that uses reliable instruments.

Random Assignment

The goal of random assignment is to establish equivalent groups for comparison. Evaluation plans should describe the unit that will be assigned (such as individuals, schools, or clinics); how the eligible sample for assignment will be identified and selected; and the process by which individuals or groups will be assigned to program or comparison conditions. Plans should also specify when random assignment will occur with respect to acquiring study consent, administration of baseline surveys, and the start of the program.

Selecting a Comparison Group

For QEDs, the characteristics of the comparison group target sample are important to understand for assessing the probability that it can provide a valid estimate of the outcomes for the treatment group in absence of the program (that is, a valid counterfactual). Providing detailed information about the demographics, location, and experiences of this population is crucial. Equally important is a discussion of how the comparison group will be formed and, in particular, whether this process encourages selection of equivalent types of individuals into the two study groups on factors that are not easily observed—such as motivation or ability to seek program services and the need for these services. Finally, the plan should document any available evidence that similar strategies have worked in the past to gain the cooperation of this target comparison group.

Minimizing Bias by Maximizing Design

Strong designs require remarkably careful implementation to obtain statistically valid findings. Thinking through the potential consequences of design decisions a priori—particularly regarding sample enrollment and tracking—and documenting those decisions in the plan are important for allowing plan reviewers to assess the probability that the integrity of the design will be maintained. For example, decisions regarding the timing and process for consent could result in high overall or differential response rates across the study groups, which might also result in nonequivalent study groups. Self-selection or nonrandom allocation of program youth or program staff in an RCT can introduce bias that could also result in nonequivalent program and comparison youth. Study plans should, therefore, describe deliberate approaches that evaluators and program staff will take to gather consent and deliver the program in a way that will maintain equivalent factors between the two groups except for the offer of programming to one of them.

As the study progresses, achieving high response rates is critical not only for maintaining the power to detect impacts, but also to ensure that overall and differential attrition do not fall below critical thresholds established by the HHS evidence review. This is particularly important for RCTs, in which strength lies in the fact that those originally randomized do not differ along any dimension except for the offer of the program. Evaluation plans should describe strategies to track the sample and achieve high response rates across both groups. It is also extremely important to document plans to collect data from individuals or sites originally randomly assigned and analyze outcomes for them, regardless of whether they continue to participate in the program.

Confounding Factors

Confounding factors make it impossible to disentangle the impact of the program from other influences. A confounding factor exists when any one element of the design aligns perfectly with either the program or comparison condition. For example, if only one school is assigned to receive a school-based program or if the mode of data collection systematically differs for the treatment and comparison groups (for example, using a survey for the treatment group but gathering administrative records for the control group), the design has a confounding factor. Confounding factors can also be related to the mode of program delivery and the method of data collection. Program implementation plans for the treatment and comparison groups (if applicable) and data collection plans should be well documented so that they can be assessed for factors that align perfectly with either the program or comparison condition.

Outcome Data Collection

Evaluation plans should identify the outcome data to be collected and lay out a schedule for collecting them, relative to other major program implementation milestones and in concert with the theory of change. It is important for outcome data collection to occur at critical points—before the program begins and again when the program is expected to have affected key outcomes. For QEDs, establishing baseline equivalence of the analysis sample on key outcomes is critical. To meet HHS evidence standards, the analytic sample has to be equivalent on age or grade level, race/ethnicity, gender, and, for samples age 14 or older, at least one measure of sexual behavior, such as sexual initiation.

Employing established and validated measures will yield the most convincing results. Therefore, study plans should document the reliability of the measures used for key outcomes. If new measures are being developed, the plan should describe efforts to pilot and assess reliabilities and functionality, and allow time to make changes as necessary.

Providing the Program to the Comparison Group

Promising the program to the comparison group in the future is a common sample recruitment strategy. If the program will be offered to the comparison group at a later time, the evaluation plan should demonstrate that these services will be provided after all data collection activities occur, so that all tests used to measure program impacts will maintain the integrity of the program and comparison contrast built into the study design.

Describe Plans to Measure Implementation and Context

Effectiveness evaluation plans tend to focus on the impact study design and stop short of describing additional data collection activities that are important for assessing whether results will be understood and lend themselves to interpretation. Impact estimates are only partially informative if what happens inside and outside the "black box" of the program model is unknown. Collecting data on implementation fidelity and context scientifically and independently can later support hypotheses about why program impacts may or may not have been observed and suggest reasonable next steps for modifications to the program model or future research.

Measuring Fidelity and Describing Context

Fidelity refers to whether a program is implemented as intended; plans to measure fidelity should identify the specific elements of the program model that will be assessed and the specific data elements that will be collected. When planning this component of the evaluation, researchers should also assess the experiences of the comparison group, particularly when an alternative program is being provided as the counterfactual condition. Assessing fidelity should include gathering information from various informants close to the provision of programming about factors that they believe may have facilitated or impeded program implementation.

Context refers to factors that might prevent the program from being implemented as intended and demonstrating impacts. These factors include characteristics of the organizations and communities within which the program is being provided, such as the availability of related services and activities and participation in them by members of the program and/or the comparison group. The elements of context to be assessed will understandably vary, partly depending on (1) how and where the intervention is provided and (2) the existing services and activities.

Proceeding Scientifically and Independently

In addition to describing the types of data related to fidelity and context that will be collected, evaluation plans should describe data collection procedures, including who will collect each data element and how. If data will be collected from subsamples, the plan should demonstrate that those decisions will be made objectively and transparently.

Although it may be less resource intensive for program staff to lead these data collection efforts, doing so reduces the necessary objectivity and leaves the findings open to critique. Plans should, therefore, clarify that evaluators, and not program staff, will lead the instrument development and data collection. In some instances, it might be appropriate for program staff to collect objective data and provide it to the evaluator, such as youth attendance records, documentation of staff qualifications, agendas from staff training sessions, lesson plans from intervention sessions, or curriculum materials.

Conclusion

Organizations increasingly make funding decisions based on prior evidence of program effectiveness, and they require that their funded effectiveness evaluations meet specific scientific standards. Currently, HHS provides funding to replicate evidence-based teenage pregnancy prevention models through multiple grant programs, and it funds evaluations to acquire new high-quality evidence on promising teenage pregnancy prevention models.

In a related effort in the U.S. Department of Education (ED), higher levels of funding for Investing in Innovation grants were available for programs that had high-quality evidence of effectiveness. Here, too, grantees are expected to conduct effectiveness evaluations that will ultimately meet the ED's scientific standards (those of the What Works Clearinghouse).

This focus on planning and maintaining high-quality effectiveness evaluations that will meet established scientific standards has resulted in a shift within federal agencies—and beyond—from using standards to assess completed evaluations toward using standards to assess the probability that evaluation plans will result in a high-quality final study. Evaluators will therefore increasingly need a useful tool—such as the material presented in this brief—when structuring evaluation plans that will convince decision makers that their design is rigorous and that the evidence will be useful.

This brief was written by Susan Zief, Juliette Henke, Jean Knab, and Heather Zaveri from Mathematica Policy Research for the HHS Office of Adolescent Health under contract #HHSP23320095642WC, Task Order No. HHSP23337017T.

MATHEMATICA Policy Research