Teen Pregnancy Prevention (TPP) Replication Study

Impact Study Design Report

October, 2015

Contract No: HHSP23320095624WC
Order No. HHSP23337011T

Prepared for:
Lisa Trivits, Ph.D.
U.S. Department of Health and Human Services
200 Independence Ave., SW
Room 404E
Washington, DC 20201

Amy Farb, Ph.D.
Office of Adolescent Health
Office of the Assistant Secretary for Health
U.S. Department of Health and Human Services
1101 Wootton Parkway, Suite 700
Rockville, MD 20852

Submitted by:
Abt Associates Inc.
55 Wheeler Street
Cambridge, MA 02138

Table of Contents

Introduction

The Teen Pregnancy Prevention Replication Study offers a unique and exciting opportunity to learn from the significant investment made in evidence-based teen pregnancy prevention programs through the Teen Pregnancy Prevention (TPP) Program, administered by the Office of Adolescent Health (OAH). The goal of the evaluation—to contribute important information to the research base on teen pregnancy prevention programs—will be accomplished through a series of rigorous experimental design evaluations of a set of evidence-based programs that are being replicated by grantees under Tier 1 of the TPP Program. These studies will investigate whether evidence-based programs, when replicated with fidelity by grantees, produce behavioral impacts similar to those demonstrated in the original studies, and will determine whether these impacts are sustained over a longer period than these earlier studies examined. The evaluation comprises two linked studies: a study of the impacts of three program models on youth who participate (the impact study); and a study of the contexts in which the programs are implemented, the extent to which they are implemented with fidelity to the original model, and the challenges faced in implementing them (the implementation study).

The design of the evaluation offers an opportunity to move beyond the question of the impact of a single replication of a program model to look at variation in impacts for program models implemented in different settings and/or with different populations. A comprehensive implementation study will allow us to examine the relationships between variation in impacts and program implementation. In addition, it will provide critical information about the contexts in which evidence-based programs are put in place, the challenges encountered, and the aspects of program implementation that may be associated with program impacts.

This report focuses on our design for the impact study. A companion report describes the implementation study. The report begins with an overview of the policy and research context for the evaluation. The chapter that follows describes the objectives, activities, and decisions of the feasibility and design contract that preceded the current evaluation contract and that laid the foundation for the final evaluation design described here. The remaining chapters present the design of the impact study. The appendices to this report contain the site-specific evaluation designs developed for each of the nine grantees selected for the evaluation.

1. Background

A major priority for HHS is finding ways to reduce adolescent risky sexual activity, sexually transmitted diseases, and pregnancies/births. A key strategy to achieve this goal is through investing in evidence-based pregnancy prevention strategies and targeting populations at highest risk for teen pregnancy. The Teen Pregnancy Prevention Initiative, which includes programs funded and/or administered by different offices within HHS, underscores the cross-cutting nature of the problem and the strategies to address it.

OAH's Teen Pregnancy Prevention Program is intended to address high rates of teenage pregnancy by (1) replicating evidence-based prevention models, and (2) testing innovative strategies. The program's funding is structured to maximize investments in programs that have been shown to be effective, but at the same time provide support for research and demonstration grants that provide an opportunity to add to the existing knowledge base.

Evaluation Efforts within the Teen Pregnancy Prevention Initiative

The Teen Pregnancy Prevention Program uses a 'tiered' approach to funding a range of programs: Tier 1 (Replication) funds were allocated for replication of programs that have demonstrated effectiveness through rigorous evaluation; and Tier 2 (Research and Demonstration) funds were allocated to programs that are partially supported by evidence (either because they build on elements of evidence-based programs or have preliminary evidence of effectiveness but have not yet been rigorously tested). With this strategy, the Federal government balanced an emphasis on evidence-based programs with the recognition that support for innovation is also important.

To ensure that the investment would add significantly to the sparse amount of strong evidence in the field, funding for both types of grantees was accompanied by requirements for evaluation activities. First, all grantees funded under Tier 1 and Tier 2 are required to conduct a careful study of the fidelity of their implementations of the program models they have chosen. Second, all grantees are required to report performance measures for participants in their programs. In addition, all Tier 2 grantees and a subset of Tier 1 grantees (those with the largest funding awards) are required, as a condition of funding, to conduct a rigorous evaluation using an independent evaluator and estimate the intervention's effects on pregnancy and sexual risk behaviors, the reduction of which is the primary goal of the initiative.

In addition to the grantee-led evaluation efforts, HHS has funded complementary evaluation activities conducted by the federal government. One of these federally-managed evaluations, the Pregnancy Prevention Approaches (PPA) study, includes evaluations of seven program models, six of which are research and demonstration grants funded through the Teen Pregnancy Prevention Program.[1] The PPA evaluation will provide evidence about the effectiveness of new and untested program models in preventing teen pregnancy and sexual risk behavior.

[1] Other HHS efforts to address teen pregnancy include the Personal Responsibility Education Innovative Strategies (PREIS) programs, along with the State Personal Responsibility Education Program (PREP).

2. Feasibility and Design Study for the New Federal Evaluation

In addition to all of these evaluation activities, the TPP legislation included funding for a new Federal evaluation and, in September 2010, a contract was awarded to Abt Associates to examine options for the focus of the evaluation, develop design parameters, recommend an overall evaluation approach and identify and recruit grantees for the new evaluation.

In the face of the wealth of research on this topic that the two sets of evaluation efforts (grantee-level and Federal evaluation efforts) represented, the major question facing the new Federal evaluation was one of direction. Should a new effort add to the multiplicity of planned studies of individual programs, adding an additional 8-10 programs to PPA's seven programs and the 40 grantee-level evaluations (OAH's TPP and ACF/FYSB's PREIS grantees)? Or should a new evaluation effort focus on different questions of policy interest?

Choosing a Direction/Focus for the New Federal Evaluation

The choice of a direction for the new Federal evaluation was governed by many considerations: the policy interests and priorities of the Federal partners; gaps in the existing research and priorities among them; and the ways in which the funded activities of grantees might be used to address their policy and research priorities.

Across a wide range of research fields, there is increasing recognition of the tension that exists between supporting and extending the use of evidence-based practices and encouraging innovation that will strengthen or replace them. The tiered structure of OAH's TPP Program acknowledges the importance of each of these strategies. Through its funding for Tier 2 programs, the initiative asks the question:

- What innovative approaches (e.g. adapting evidence-based program models for use with special populations; strengthening evidence-based programs by adding components; testing new program models) are effective in reducing teen pregnancies and births to teens?

The funding for Tier 1 programs has the potential to address the question:

- Do replications of evidence-based program models produce impacts similar to those originally demonstrated, as well as effects on teen pregnancy and births to teens?

The PPA evaluation is designed to address the first question. The required evaluations of all Tier 2 and PREIS grantees will also address it. By contrast, only the largest grants to Tier 1, 16 out of 75 grantees, carry a requirement for rigorous evaluation. These facts suggested that the new evaluation could supplement these existing evaluations, first, by focusing on replication of evidence-based programs and second, by identifying a strategy that moved beyond the evaluation of individual replications.

For the TPP Replication Study, OAH has chosen to focus on addressing the second question by selecting a small number of program models from those being replicated and, within each model, selecting multiple replications. The advantage of this approach is that it allows for pooling of

data as well as representation of variation in community context or populations targeted.[2] The strategy has the disadvantage of constraining the number of program models that can be included, but the advantages outweigh the disadvantages. Given the likely resource constraints, we recommended that the evaluation include three program models, each with at least three replications.

Selecting Program Models and Replication Sites

The two steps in the evaluation design process were selection of program models; and, within each program model, selection of at least three replication sites.

Selecting Program Models

HHS conducted a pregnancy prevention research review of more than 1,000 studies and found 28 program models that met effectiveness criteria that included strength of study design, outcomes related to reduction of sexual risk behavior.[3] Only proposals to replicate one or more of these 28 program models were considered for funding under Tier 1 of the OAH grant program. A majority (24 of 28) of the evidence-based program models on the HHS Pregnancy Prevention Evidence Review are being replicated by OAH Tier 1 grantees.[4] Since only a small number of these could be included in the new Federal evaluation, HHS staff needed to weigh the relative policy importance of the different program models. A program model might be considered of policy importance if it is currently widely used, if it addresses a population of interest, or if it is being implemented in a new setting. For the evaluation, any program model selected needed to have at least five replications, since it was unlikely that all of the five would be able to meet the requirements imposed by participation in a rigorous Federal evaluation. Finally, to the extent possible, the program models chosen, as a group, should reflect variation in their approaches to teen pregnancy prevention.

A review of the successful 2010 grant applications identified nine program models with five or more replications. After discussions with HHS, we eliminated from the list the CAS-Carrera model, which has nine replications. The program itself is remarkable in the breadth of its approach and in its duration and intensity. These qualities make it notably more expensive than any other program and thus able to serve only a small number of youth over a period of four years. These factors make widespread adoption unlikely, limiting its policy relevance. In addition, the small number of youth served at any one time presents a challenge for a rigorous evaluation that requires a sample large enough to detect impacts on sexual behavior outcomes and teen pregnancy.

---

[2] Note that, while pooling data would allow for comparison of differences in impact for different populations or ethnic groups, the same statistical analysis would probably not be possible for different settings, since even pooled data would probably not provide a sufficient number of settings. It would, however, be possible to look at the contribution that "setting" makes to variation in outcomes, a less rigorous, but informative analysis.

[3] The list has been revised to incorporate additional studies that were available after the initial list was developed. There are currently 31 evidence-based programs on the list. The review criteria can be found on the OAH website: http://www.hhs.gov/ash/oah/oah-initiatives/tpp/eb-programs-review-v2.pdf.

[4] This is true of the original funding decisions, although there was some shifting from one program model to another as grantees began to investigate the availability and cost of training materials and sessions.

Of the remaining eight program models, the Teen Outreach Program (TOP) had the largest number of replications and funding resources allocated. However, of the seventeen grantees proposed to replicate TOP, seven were required to conduct a rigorous evaluation of the program. In view of the number of evaluations already planned, our recommendation was to eliminate TOP from consideration for the new evaluation and focus resources *primarily* on program models that would not otherwise undergo a rigorous independent evaluation.

From the seven remaining programs, HHS selected three that, as a group, reflect variation in program focus, service delivery strategy and populations targeted. Appendix A provides an illustration of the flow of program models and grantees into the study. ***Safer Sex*** (*SSI)* is a clinic-based program that targets female adolescents ages 14-19 who are sexually active—a group that is at very high risk for teen pregnancy.[5] ***Reducing the Risk*** (*RtR*), by contrast, is a curriculum-based program, widely used in classroom settings (as well as some community-based settings) with students, a majority of whom are not yet sexually active, even in high risk communities, such as those targeted by the TPP Program. ***¡Cuidate!*** falls between the two extremes, geared toward Latino adolescents 13-19 who are at high risk for HIV/AIDS, not all of whom are sexually active at the time they receive the program. The program is widely delivered in school and community-based settings.

The three programs differ in their target population, strategies for delivering service, and the duration and intensity of the service provided. *SSI* serves female youth only whereas *¡Cuidate!* and *RtR* serve both males and females. *SSI* provides one-on-one counseling to individual female youth in four sessions spread over six months; the *SSI* curriculum mandates a set of topics to be covered in the first session and provides minimal scripting for all of the sessions. *¡Cuidate!* includes six sessions which can be delivered over two days or over one to six weeks, to small groups of 10-12 youth. The program provides topics for each session and culturally-appropriate materials. *RtR* has 16 highly-scripted sessions for groups that can range in size from 15 to 30 or larger. The program may be delivered over a semester or a shorter period of time, depending on the length of time allocated for the class.

The program models and their logic models are described in more detail below.

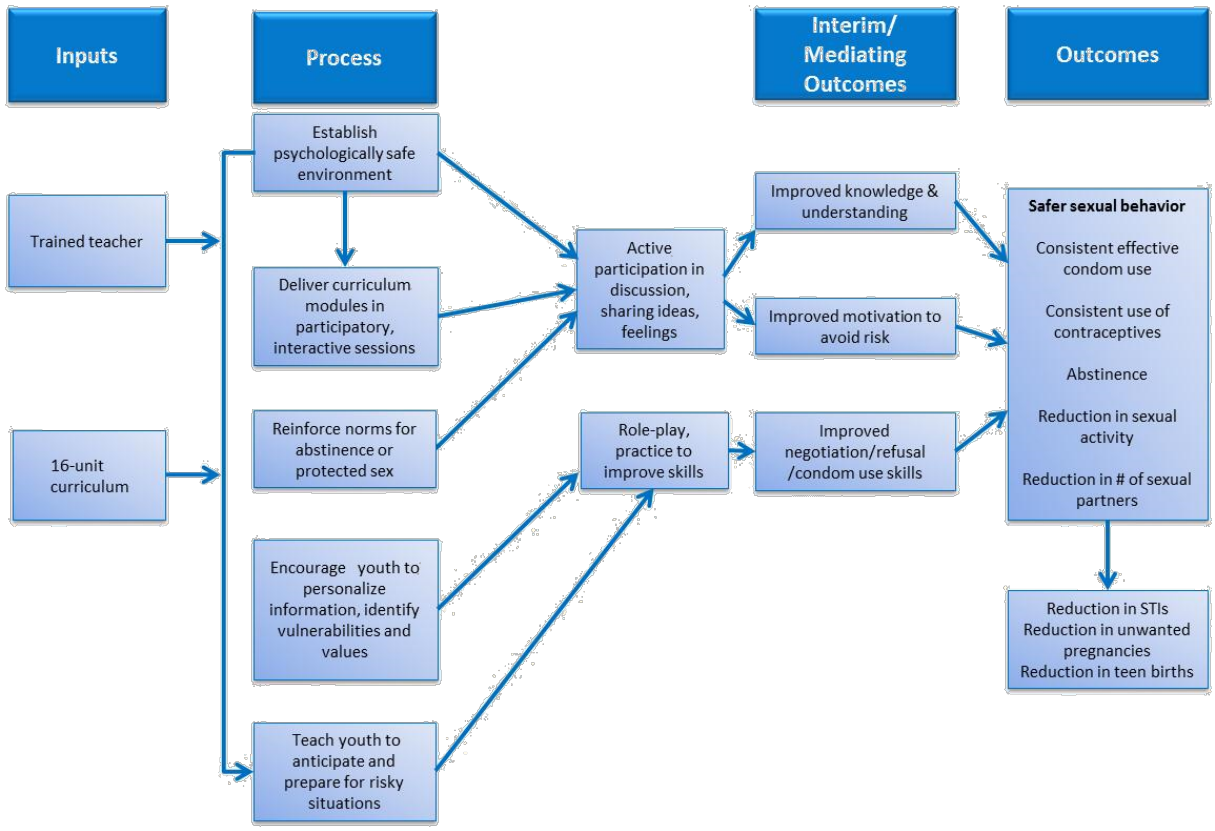
Reducing the Risk is a sexual health curriculum designed for use in high school classrooms, which can also be implemented in other community settings where youth receive services. The program's overarching goal is to prevent pregnancy and STDs among high-school-age adolescents, by changing four sexual behaviors directly related to the goal: amount of sexual intercourse; initiation of sexual intercourse; use of condoms; and use of contraceptives.

Exhibit 1 shows the program elements, the intended outcomes and the pathways by which the program seeks to achieve these outcomes. A trained teacher or health educator delivers the sixteen 45-minute units of *Reducing the Risk* in a classroom or other setting. The first objective for the teachers is to create an environment of mutual trust in which youth can speak freely about their attitudes, feelings, values and perceptions. Within that atmosphere of trust, the teacher delivers the 16 modules in a planned sequence. As part of every module, the teacher reinforces the norms of abstinence and protected sex. The sessions are interactive and encourage active

[5] The original *SSI* included females ages 14-23.

participation by students. Youth are encouraged to personalize the information, identify their own vulnerabilities and examine their personal values. The sessions repeatedly offer opportunities for youth to anticipate and prepare for situations in which they may be pressured to have unwanted or unsafe sex, and to practice the skills they need to deal with these and similar situations. Taken together, the sessions are intended to increase students' knowledge and understanding of sexual health issues, correct unfounded beliefs, develop more positive values, attitudes and intentions with respect to abstinence and unprotected sex, and develop their communication, negotiation and refusal skills. These interim outcomes mediate the behavioral outcomes that the program seeks to achieve: abstinence from sex, delay in initiating sex, and correct and consistent use of condoms and birth control for those who are sexually active. Prevention of or reduction in sexually risky behavior is ultimately expected to reduce rates of pregnancy and births, as well as STDs among teens.

Exhibit 1: Reducing the Risk Logic Model

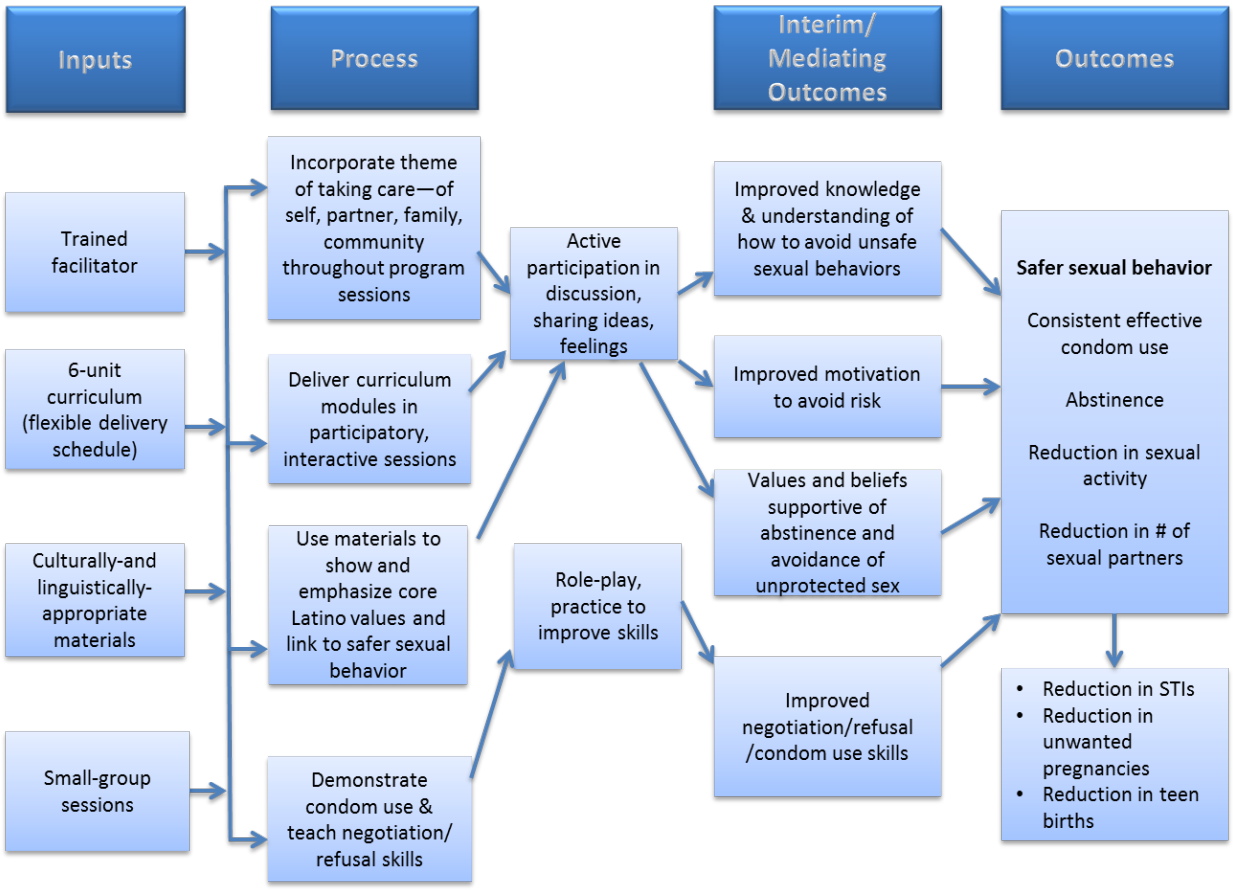


¡Cuidate! is adapted from the *Be Proud! Be Responsible!* curriculum and culturally tailored for Latino youth. It aims to reduce HIV risk and unintended pregnancies by affecting sexual behaviors such as frequency of first intercourse, number of partners, and condom use. The program integrates cultural beliefs and attitudes in the Latino community (such as familialism and machismo) to communicate the importance of risk-reduction strategies and to increase knowledge and self-efficacy skills. The program consists of six modules of 60 minutes each delivered over a two-day period (or longer) in small groups of 10-12 youth ages 14-19. The modules are led by trained adult facilitators who are bilingual in English and Spanish. The program has been implemented in an after-school setting on consecutive weekends, but can be

delivered in other settings, such as community-based organizations and during the school day, as well as on schedules that vary from the original.

Exhibit 2 shows the program elements, the intended outcomes and the pathways by which the program seeks to achieve these outcomes. A trained facilitator leads six hour-long sessions with small groups of teens in a school or other setting, using culturally-appropriate materials. The curriculum modules are delivered in participatory, interactive sessions. Each session weaves in the theme of Taking Care – of oneself, one's partner, family and community. The materials used in the sessions emphasize core Latino values and feelings, and link them to safer sexual behavior. The facilitator uses a condom model to demonstrate correct use, and teaches negotiation and refusal skills. Through active participation in discussions, sharing ideas and feelings and role-playing situations in which they may be pressured to have unwanted or unsafe sex, participants increase their understanding of sexual risks and safe sexual practices and their motivation to avoid these risks. Through repeated role-play they acquire the skills they need to deal with unwanted pressures and risky situations, refuse unsafe sex and negotiate safe sex, and use condoms correctly. These outcomes mediate the behavioral outcomes that the program seeks to achieve: abstinence from sex, delay in initiating sex, reduced sexual activity, and correct and consistent use of condoms and birth control for those who are sexually active. Prevention of or reduction in sexually risky behavior is ultimately expected to result in reduction in the rates of pregnancy and births, as well as STDs among teens.

Exhibit 2: Logic Model for ¡Cuídate!



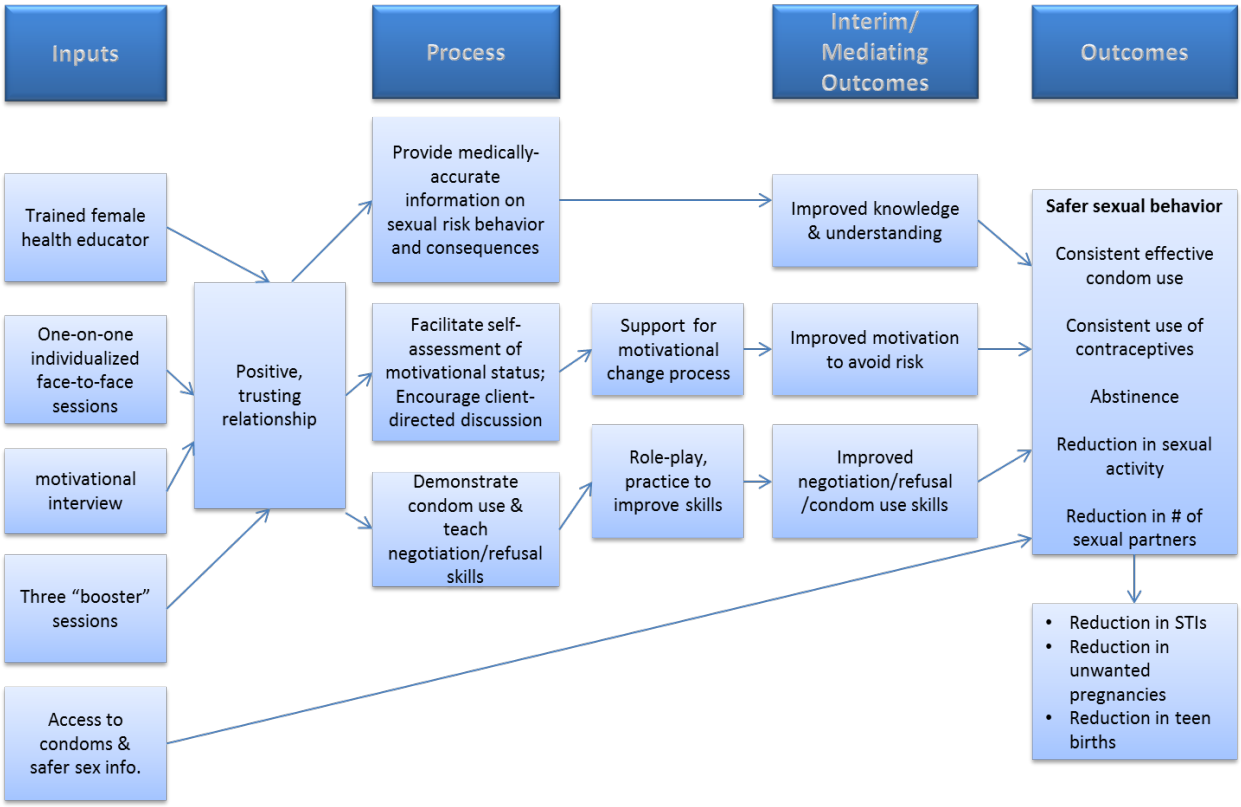
***Safer Sex** (SSI)* is a clinic-based program designed for young women ages 14-23 who are at high risk for sexually transmitted infections (STIs) and unintended pregnancy. The goals of the program are to reduce sexual risk behaviors, increase condom use, and prevent the incidence or recurrence of STIs among sexually active young females. To achieve these goals, the program seeks to capture the attention of young females, deliver information about how to engage in safer sex, and promote attitudinal and behavior changes.

SSI adopts a motivational interviewing framework in which a health educator delivers the intervention in one-on-one, face-to-face sessions with the young female. Using motivational interviewing techniques, health educators tailor program messages to each individual's unique circumstances and needs. The intervention has two versions: a Pre-Contemplation Stage Module, which emphasizes delivering information and obtaining feedback about safer sex behaviors; and a Contemplation Stage Module, which emphasizes education, skills, self-efficacy and self-esteem. The choice of which version to use is made by the health educator on the basis of the client's self-assessment on the Wheel of Change, their subsequent discussion and the health educator's own assessment of the client. Using a videotape to introduce information about condom use, the Wheel of Change for self-assessment and reflection, and a motivational interviewing strategy to encourage participant-directed discussion, the health educator guides the 50-60 minute session through a sequence of topics and allows time for role-plays, questions, and feedback on the session. Three subsequent booster sessions, delivered one, three and six months after the initial session, can vary in length from 10-20 minutes, depending on the needs and interest of the client, and are used to review information, assess progress and provide additional information and practice, if needed. Participants are offered condoms and informational materials.

Exhibit 3 shows the *SSI* logic model including program elements, the intended outcomes and the pathways by which the program seeks to achieve these outcomes.

The program's theory of action suggests that a trained health educator, using motivational interviewing strategies during one-on-one, face-to-face, individualized counseling sessions and subsequent booster sessions will establish a positive and trusting relationship with the client. In this context, the educator provides medically-accurate information, facilitates self-assessment, encourages a client-directed discussion about risky sexual behavior and relationship issues, demonstrates condom use and teaches negotiation skills. Through question and answer, discussion and role-play, and the educator's support for behavioral change, the client gradually shows improved knowledge and understanding of sexual risk behavior and its consequences, is more motivated to avoid risk and more able to negotiate safe sex and refuse unsafe sex. Greater understanding of the consequences of risky sexual behavior, combined with improved motivation to avoid risk and better negotiation skills are mediating outcomes that lead to the outcomes of interest: namely safer sexual behavior (consistent, effective use of condoms and other contraceptives, abstaining from or reducing sexual activity or reducing the number of sexual partners). Ultimately, those behaviors will lead to reductions in STIs, teen pregnancies and teen births.

Exhibit 3: Safer Sex Logic Model



Selecting Replication Sites

Each of the program models selected is being replicated by at least four grantees.[6] Our evaluation design called for selection of at least three replications of each model. Complicating the selection of replications was the fact that most of them were not designed with the requirements of a rigorous experimental evaluation in mind. The grant announcement specified as a condition that, if selected, the grantee must agree to participate in the Federal evaluation, but did not spell out what that might mean. In some cases, schools or other partners had signed agreements with grantees to implement the program but had no such agreement about evaluation. Sometimes these agreements could be renegotiated but, in other cases, districts were unwilling or unable to participate in research activities and ready to decline the program if it meant participating in an evaluation. In other cases, grantees were struggling to reach agreement with school districts to implement the program and it was unclear whether they would be successful, even without the added burden of an evaluation. In some replications, the control condition (the services that individuals would receive if they were not assigned to the program) was not sufficiently different from the program model tested to allow for a strong test of the model. In almost every case, it was clear that it would be necessary to build the evaluation sample over a two-year period, to achieve the necessary sample sizes for an experimental study.

[6] At the time of model selection, all three of the models had been selected by at least five grantees, but some grantees changed their selection in the course of the first year.

Working with grantees intensively over a period of more than a year, we were able to identify nine grantees willing and able to participate fully in the evaluation and meet all of its requirements. Appendix A illustrates the selection of program models and grantees into the study. Exhibit 4 summarizes the characteristics of each program model and its replications. The description of each program model shows the characteristics of the original model tested. The description of the replications includes any OAH approved adaptations of the model, where this is relevant.

Exhibit 4. Key Features of Program Replications in the Evaluation, by Program Model and Replication Site

Program Model, Grantee	Program Description	Study Location	Target population: Age	Target population: Demographics (from proposal description)	Program Duration and Intensity	Program Setting	Program Delivered By
Original Evaluation Study							
Reducing the Risk[7]	Sexual health and risk prevention curriculum delivered to groups in schools or community settings	13 high schools throughout CA (46 classes)	High school students, mixed gender	62% white, 20% Hispanic, 9% Asian, 2% African American, 2% Native American	16 45-minute sessions, which can be doubled-up.	High schools	Teachers
Grantees Replicating the Program							
Better Family Life		St. Louis City, MO, St. Louis County, MO and St.Clair County, IL	9th graders (with small numbers of 10th and 11th graders.	98% African American; low SES (75% eligible for free/reduced-price lunch in St. Louis City); high risk for teen births and STIs	16 sessions delivered over 8 to 16 weeks, depending on school schedule	Non-core classes in 6 high schools	Health educators trained and employed by BFL
LifeWorks		Austin, TX	9th graders (with small numbers of 10th and 11th graders)	75% minority youth, below poverty level; high rates of teen pregnancy; high rate of STIs	16 sessions delivered over 8 weeks	Health classes in 5 high schools	Health educators trained and employed by Planned Parenthood (grant partner)
San Diego Youth Services		San Diego County, CA	9th graders (one school with 8th graders)	9th and 10th grade students in the county in schools identified as "teen pregnancy hotspots" by the state	16 sessions delivered over 8-16 weeks depending on school schedule	PE/health classes in 7 high schools	Health educators trained and employed by 5 agency grant partners

[7] Kirby, D., Barth, R. P., Leland, N., & Fetro, J. V. (1991). Reducing the risk: Impact of a new curriculum on sexual risk-taking. *Family Planning Perspectives, 23*(6), 253–263. This study found no effects after 6 months, but after 18 months, female, but not male, adolescents in the program who were sexually inexperienced at baseline were significantly less likely to report having had unprotected sex. No significant effects were found on sexual initiation, recent sexual activity, or pregnancy.

Program Model, Grantee	Program Description	Study Location	Target population: Age	Target population: Demographics (from proposal description)	Program Duration and Intensity	Program Setting	Program Delivered By
Original Evaluation Study							
¡Cuídate![8]	HIV/AIDs prevention program for small groups with emphasis on Latino cultural values	Saturday program serving neighborhoods in northeast Philadelphia	Adolescents 13-18 years of age, mixed gender	All Latino, 85% Puerto Rican	Six one-hour sessions that can be delivered over 2 days to six weeks	After-school programs or community-based organizations	Trained facilitators
Grantees Replicating the Program							
Touchstone Behavioral Health		Phoenix, AZ	8th graders	61% Hispanic, 29% white, 7% African American; 18.5% below Federal poverty line	Approved adaptation added 1 session on pregnancy prevention. 7 modules delivered over 3 weeks.	Non-core classes in 10 K-8 elementary or intermediate schools	Facilitators hired and trained by TBH
La Alianza Hispana		Boston, Chelsea and Lawrence, MA	9thgraders (some 10th and 11th graders)	62-78% Hispanic, 9-20% white, .4-25% African American; 68-88% free/reduced-price lunch	Six sessions varying by school from nine 45-minute sessions over 3 weeks to three 2-hour sessions in one week.	Non-core classes in 2 high schools, after -school program in 1 high school	Facilitators hired and trained by LAH
Community Action Partnership of San Luis Obispo County		SLO County, CA	10th graders	29-47% Hispanic, 47-64% white, 1-3% African American; 35-50% free/reduced-price lunch	Approved adaptation added 2 sessions on STIs and pregnancy prevention. Eight sessions over 8 weeks	Pullout sessions during school day in 3 high schools	Facilitators hired and trained by CAPSLO

[8] Villarruel, A. M., Jemmott, J. B., & Jemmott, L. S. A randomized controlled trial testing an HIV prevention intervention for Latino youth. (2006). *Archives of Pediatrics & Adolescent Medicine, 160(*8), 772–777. This study found that adolescents in the program were significantly less likely to report having had sexual intercourse and multiple partners in the previous 3 months; they reported significantly fewer days of unprotected sex and more consistent condom use. No significant effects were found on condom use at last sex or the proportion of days of sexual intercourse that were condom protected.

Program Model, Grantee	Program Description	Study Location	Target population: Age	Target population: Demographics (from proposal description)	Program Duration and Intensity	Program Setting	Program Delivered By
Original Evaluation Study							
Safer Sex[9]	HIV/AIDS Prevention program for high-risk females younger than 24	Urban children's hospital; adolescent clinic	Adolescent females who are not pregnant	49% African American, 18% Hispanic, 14% Non-Hispanic, White; all sought treatment for an STI at clinic	Initial one-hour face-to-face session with three 30-minute booster sessions over six month period	Health clinics	Female health educator
Grantees Replicating the Model							
Planned Parenthood of Greater Orlando		Orange County and adjacent counties, FL	Sexually active females ages 15-19, who are not pregnant	72% white, 21% African American, 25% Hispanic, 5% Asian; 41% of children living in economic hardship; high rates of STIs		Two PPGO reproductive health clinics in Orlando	Health educators trained and hired by PPGO
Knox County Health Department		Knox County and adjacent counties, TN	Sexually-active females ages 13-19 who are not pregnant	89% white, 9% black, 19% females 15-19 are Latina; poverty rates up to 34% for children under 18; many teens from high risk situations; serve children in state custody		16 reproductive health, adolescent health clinics	Health educators trained and hired by Knox County Health Department and grant partners
Hennepin County Human Services and Public Health Department		Hennepin County, MN	Sexually-active females ages 13-19 who are not pregnant	32% African American, 10% Latino, 46% Caucasian; large disparities in family income by race/ethnicity; sites selected have teen birth rates approaching or exceeding the national teen birth rate		19 reproductive health, adolescent health, school-based health clinics	Health educators trained and hired by Hennepin County and grant partners

[9] Shrier L.A., Ancheta R., Goodman E., Chiou V.M., Lyden M.R., & Emans S.J. (2001). Randomized controlled trial of a safer sex intervention for high-risk adolescent girls. *Archives of Pediatrics & Adolescent Medicine, 155(*1), 73-9. This study found no effects one month after the program, but six months after the program, adolescents who participated in the program were significantly less like to report having had another sexual partner, aside from their main partner, in the prior six months.

3. Design of the Impact Study

In this chapter, we provide an overview of the design for the impact study (the implementation study is described in a companion report) that includes experimental tests of three separate program models including a total of nine separate replications.

Overview

The impact study will estimate the effects of three replications of each of three evidence-based program models and, for each program model, will address the following research questions:

1. What are the average program impacts on teen pregnancies/births, sexually transmitted diseases (STDs), and/or sexual activity (e.g., contraceptive use, number of partners, sexual initiation, etc.)?

    a. What are the average program impacts on intermediate outcomes such as knowledge of and attitudes towards sexual risk behavior, motivation to avoid risk behavior and negotiation skills?

    b. Do the average impacts on any of the primary or intermediate outcomes differ for certain subgroups (e.g., gender, age, ethnicity, sexual experience at baseline)?

2. Is there variation in average impacts across the sites replicating a specific program model?

In each of the nine replication sites, youth in both treatment and control groups will be surveyed at three time-points, with the schedule for data collection differing slightly by program model. Survey measures designed for use in all of the HHS Federal evaluations in the TPP Initiative will be converted for web-based audio computer-assisted self-interview (ACASI) administration. Data will be pooled across sites within each of the three program models to determine the average impact of the programs, and then analyzed to determine whether the impact varies across sites within a program model. The impact study will produce impact estimates for the following behavioral outcome domains: (1) sexual risk behavior, (2) incidence of STDs, and (3) incidence of teen pregnancy. It will also produce impact estimates for the following intermediate outcome domains: (1) knowledge of and attitudes towards the risks of sexual activity, (2) intentions to avoid risky sexual behavior, and (3) skills in negotiating over condoms and birth control, and whether to engage in sex. In addition to estimating impacts in these domains, an exploratory component of the impact analysis will consider mediation (the pathways through which programs achieve realized impacts) and dosage (the time each youth spends in program activities).

Common Theory of Action for the Three Programs

While the impact study will evaluate the impacts of three separate program models, with program-specific logic models, they all share a common theory of action which informs our approach to the evaluation. The theory of action underlying the three interventions suggests that if the interventions are to be effective, we would expect their implementation to produce the following chain of events:

1. The program conducts planned activities, is implemented with fidelity and youth are responsive and engaged (verified by our implementation study).

2. Participation in the program leads to change in a set of mediating outcomes:

a. ***Increased knowledge*** (more accurate knowledge and understanding of the risks associated with sexual activity—what STDs are and how they are transmitted, pregnancy risk).

b. ***Protective attitudes*** (intention to delay sex, belief that condoms can protect against risk, intention to use protection (if sexually active)).

c. ***Motivation to avoid sexual risk behavior.***

d. ***Better skills*** (condom negotiation; refusal skills; relationship skills).

3. These mediating outcomes lead to changes in a set of intermediate behavioral outcomes related to ***sexual risk behavior:*** delay in initiation of sex; consistent use of contraception and condoms; reduction in number of sexual partners.

4. Decreased sexual risk behaviors ultimately lead to the longer term outcomes of interest:

a. ***Decreased rate of teen pregnancy*** and/or births

b. ***Decreased incidence of Sexually Transmitted Diseases*** (STDs)

While the programs share this common general theory of how intervention activities lead to improved outcomes, the outcomes that will be most sensitive to the specific program activities are likely to vary across programs. For example, because all individuals enrolled in *SSI* are sexually active at the time of enrollment, this program cannot affect sexual debut, but it may reduce teen pregnancy and STDs through a reduction in other sexual risk behaviors.

This theory of action leads us to a study that encompasses all of the mediating, intermediate and longer-term outcomes specified in the logic model described above. All of these outcomes will be measured at baseline and at two additional points in time to estimate the effects of the three interventions on these outcomes (see later sections for more details).

Impact Study Design

This section describes the experimental design and procedures for random assignment, data collection and analysis.

The design for each program model will include two groups: (1) a treatment group and (2) a control group. Adolescents assigned to the treatment group will be invited to participate in the program; adolescents assigned to the control group will not be invited to participate in the program, but they may still receive the usual services offered in school or clinic settings.

The study will thus produce evidence on the impacts of the particular intervention being tested relative to a counterfactual in which the program model is not offered for each of the three program models. As part of our work on the TPP Feasibility and Design Study, we confirmed that the nine grantees chosen for the evaluation are operating no other programs that would be considered close substitutes for the three programs to be evaluated, that is, sexual health educational interventions that target the population of interest to the evaluation. Therefore, while study participants may receive other services from clinics, schools, and other service providers[10], the evaluation will provide a test of the "value-added" of the three intensive

[10] In some schools, for example, the mandatory health curriculum may include at least one module that covers sexual risk behavior. In those settings, youth in both the treatment and control groups will be exposed to some

programs selected for the evaluation, relative to the "standard of care" offered in the communities in which study participants reside.

Unit of Random Assignment

A key decision in the experimental design is the level of random assignment. For this evaluation, we will:

- ***Randomize individuals in each of the replications that administer the intervention "one-on-one" in a clinic-based setting, as a pull-out program in schools, or in community-based organizations***. This includes all three replications of SSI and ¡Cuidate!. We will randomize individuals because it is feasible in these settings, and because in general, randomizing individuals yields more statistical power than randomizing groups of individuals.

- ***Randomize classrooms in each of the replications that administer the intervention in school classrooms.*** This includes the three replications of *Reducing the Risk.* We elected to randomize classrooms because our experience has shown that randomizing individuals to the interventions, and thus implicitly to classrooms, may not be feasible in this setting. Randomizing classrooms is also acceptable for this study because external trained staff are delivering the intervention rather than teachers, which minimizes concerns about teachers inadvertently (and perhaps unknowingly) delivering aspects of the intervention to students in control classrooms. We will randomize classrooms *after* students are assigned to their classes to ensure the integrity of the experimental design.

Even strong random assignment designs have threats to their internal validity. For example, all social experiments face threats from noncompliance (e.g., crossovers) and study attrition. Furthermore, studies of interventions that are designed to influence how individuals interact with other individuals are particularly at risk from the bias that can result from contamination, or spillover effects from one of the experimental groups to the other. This risk could be reduced by randomizing entire schools to the treatment or control conditions. However, this would require offering the intervention on a larger scale than permitted by grant funding levels—and would require additional funding for the impact evaluation. Methods for addressing these challenges are discussed later in the chapter.

Sample Sizes

As noted earlier, a major task of the TPP Feasibility and Design Study was to identify the programs to be included in the evaluation, along with the sites that would be included for each replication. The selection of these programs and sites was driven in part by the feasibility of including programs and sites in a rigorous impact evaluation. In addition, part of the feasibility assessment involved a statistical power analysis, where the overall goal was to identify replication sites and programs that could contribute evidence *with adequate precision* to address the research questions specified for the evaluation, both for the pooled confirmatory analysis and

---

sexual health education. In some clinic settings, reproductive health services are available to members of both groups if they choose to access them. The implementation study will document the availability of such services.

for each site individually. Since statistical precision depends heavily on the size of the sample, we identified sites with samples that were deemed large enough for the experiment.[11]

Exhibit 5 illustrates the target baseline sample size and the expected sample for each of the follow-up data collection intervals. Both parent permission and informed student assent are required in order to participate in the study.[12] For *Safer Sex*, this includes all youth for whom we obtain consent. For *Reducing the Risk* and *Cuidate!*, students for whom we have parental consent but who refuse to assent on their own behalf prior to completing the baseline survey will be dropped from the sample; all others for whom we have consent (including those who are absent the day of the baseline survey) will remain in the sample. We assume that for the first and second follow-up surveys we will obtain completed survey responses for 86 and 80 percent of the study sample, respectively.[13] While this may be optimistic, our Data Collection Plan describes the significant efforts that we will take to ensure the highest response rate possible.

Exhibit 5: Sample Size Assumptions

Program Model	Baseline	Short-Term Follow-Up (86% sample retention)	Longer-Term Follow-Up (80% sample retention)
Safer Sex	2,850 (950/site)	817/site	760/site
Reducing the Risk	2,850-3,000 (950-1000/site; or 48-60 classrooms)	817-860/site	760-800/site
¡Cuidate!	2,550-2,850 (850-950/site)	731-817/site	680-800/site

Conducting Random Assignment

Our approach to conducting random assignment is designed to ensure that a rigorous evaluation can be conducted *without disrupting the normal operations of the program*. This requires separate random assignment procedures for each program. Exhibit 8 summarizes the method of conducting random assignment in each of the three program models. The process for conducting random assignment is described in more detail for each program model below.

[11] Minimum Detectable Impacts (MDIs) for specific outcomes are presented and discussed in detail in the Analytic Methods section.

[12] The Abt Associates Institutional Review Board (IRB) received a waiver of parental consent in *SSI* sites in which minors are not accompanied by a parent to the clinic. In the *SSI* clinics adolescents can consent to treatment and procedures, such as contraceptive services, pregnancy and disease testing, without parental knowledge, and therefore, Abt received a waiver to protect the privacy of the adolescent.

[13] In the second follow-up survey, we will attempt to collect data on the entire study sample, regardless of whether they responded to the first follow-up survey.

Exhibit 6: Random Assignment Approach and Target Sample Size, by Model and Replication Site

Program model	Replication site (grantee)	Unit of Assignment	RA approach	RA ratio	Target Sample size
Reducing the Risk	Better Family Life (St. Louis, MO)	Classes within schools	Random assignment before the program begins	2:1	54 classes (950+ students)
	LifeWorks (Austin, TX)	Classes within schools	Random assignment before the program begins	1:1	54 classes (950 students)
	San Diego Youth Services (San Diego County, CA)	Classes within schools	Random assignment before the program begins	varies by school	48 classes (1,000+ students)
¡Cuídate!	Touchstone Behavioral Health (Phoenix AZ)	Individuals within schools	Random assignment by gender before the program begins[14]	1:1	850 youth
	La Alianza Hispana (Greater Boston, MA)	Individuals within schools	Random assignment before the program begins	2:1	950 youth
	Community Action Partnership of San Luis Obispo (SLO County, CA)	Individuals within schools	Random assignment before the program begins	2:1	950 youth
Safer Sex	Planned Parenthood of Greater Orlando (Orange County and contiguous counties, FL)	Individuals	Rolling assignment before the intervention begins	2:1	950 youth
	Knox County Health Department (Knox County and contiguous counties, TN)	Individuals	Rolling assignment before the intervention begins	2:1	950 youth
	Hennepin County Health Department (Hennepin County, MN)	Individuals	Rolling assignment before the intervention begins	2:1	950 youth

In a clinic-based program like *SSI*, which serves adolescents one at a time, random assignment will be conducted almost immediately after a person agrees to participate. In *RtR*, random assignment will occur shortly after students are assigned to classes, and before school starts, to ensure that classroom teachers are prepared for the intervention in selected classes. School-based "pull out" programs during school, and before- and after-school programs, like some *¡Cuidate!* programs, need to fill their program slots, and random assignment is designed to help them achieve that goal. In all cases, youth will complete the baseline survey before they are told their assignment status.

Successful use of random assignment in any study requires two conditions. First, assignment to the treatment and control groups must be random (though the assignment probabilities may vary across the sample). Second, compliance with random assignment must be maintained for the duration of the study. Below, we describe our plan for conducting random assignment for each of the three programs, given the logistical considerations that these programs face in the sites chosen for the evaluation, and for ensuring that the integrity of random assignment is maintained over time in each of the nine replication sites.

[14] The program is delivered in single gender groups.

Safer Sex Intervention (SSI)

The *SSI* replication sites will use individual-level random assignment. However, the details of random assignment will depend on whether the clinic is prepared to begin delivering services on the day on which people express interest in the program.

In general, eligibility screening, informed consent and intake into the study (the baseline survey, random assignment), and the provision of the first *SSI* session (for those randomized to the treatment group) will occur on the same day the adolescent female arrives at the clinic to obtain the service she has requested. In some instances, young women may also arrive at the clinic specifically because they have been recruited for the study, and they may not necessarily be seeking clinic services. There are some circumstances in which it will not be possible to have all study and *SSI* events happen on a single visit. First, not all clinics have a full time health educator who is available to recruit potential participants as they arrive at the clinic, so they may not be prepared to provide "same day" services. In addition, potential participants who express interest in the program may have already been in the clinic for a long period of time and may not want to stay even longer to enroll in the study, take the baseline survey, be randomly assigned, and stay for an initial *SSI* session if assigned to treatment. To the extent that these challenges are faced by the *SSI* grantees selected for the Federal replication study, we are prepared to conduct random assignment slightly differently. If there is not time to complete study enrollment, baseline, random assignment, and the first *SSI* session (for those assigned to treatment), all actions associated with enrollment into the study (enrollment, baseline, random assignment, and the first *SSI* session), except the eligibility screening will take place on a subsequent day, by appointment. Under both scenarios, study participants will complete the baseline survey before random assignment.

Reducing the Risk (RtR)

Since *RtR* is a school-based program, we will use a randomized cluster design in which classes, rather than individuals, are randomly assigned to *RtR* or to a control group. Classes will be randomly assigned within each school. In each participating school, we will suggest randomizing half of the eligible classes to the treatment group. However, we plan to show flexibility in the fraction of classes assigned to the treatment group to accommodate local needs and preferences, and the program's need to reach its service targets, and we will accept an unbalanced design as long as the proportion of classes assigned to treatment is between one-third and two-thirds (since the loss of statistical power from an unbalanced design is relatively modest if the assignment rate falls within this range).

In each participating high school, classes that the school and grantee have determined to be suitable for the program will be randomly assigned either to receive *RtR* or to implement whatever curriculum would have been used in the absence of the program. The type of class selected may differ across schools, but within schools a single class type will be selected. Possible classes include Health, Advisory Period, or Social Studies. In addition, because the sample of classes will be built over a two-year period, *RtR* may be delivered in 9th or 10th grade. However, no school will implement the intervention in 9th grade in Year 1 and 10th grade in Year 2; this would risk the possibility that students assigned to the control group in Year 1 would be assigned to an *RtR* class in Year 2. In all study classes, all students whose parents provide written consent and who themselves provide written assent will be eligible to participate in the study.

Parental consent and student assent will be obtained before the treatment status of classes is made public. After students assent to the study, and before they know their treatment status,

students will be asked to complete the baseline survey in a group setting.[15] Within each participating school, classes will be randomly assigned by evaluation staff using statistical software. Classes will be randomly assigned as late as possible, but sufficiently early so that teachers can incorporate the intervention into their lesson plans.[16]

Note that, for logistical reasons, in some cases it is necessary to collect baseline data after random assignment has been conducted. While this can bias the impact estimates under some conditions (see Schochet, 2008), we believe that this will not bias the impact estimates here because the study has been designed to keep students "blinded" to their assignment until after the baseline survey has been completed. For example, we will use the same study consent form for the two groups, and we will take steps to ensure that there are no announcements either to students or parents about which classes will be providing *RtR* this school year prior to the completion of the baseline data collection.

We will randomize classes either each fall semester for two years or, in certain sites where feasible (e.g., where *RtR* could be offered in each semester), in both fall and spring semesters, until the target sample size of 48–64 classes is reached. All students in study classes in the three replication sites who have parental permission and who themselves agree to the study will be surveyed prior to the first *RtR* session. Once a student is scheduled into a control class, the student will not be able to participate in *RtR* through the end of the two-year embargo period.

¡Cuidate!

The approach to random assignment for *¡Cuidate!*, is to randomly assign individual youth, as the program is offered in "pull out" groups during, before, or after school. In the case of school-day implementation, assignment to the control group would mean that youth would not be pulled out of class to attend *¡Cuidate!*; rather, they will either remain in their classes or engage in other activities unrelated to the intervention before and after school. The length of the randomization period will depend on how many youth per group, and how many group cycles occur in a year.

Eligibility for the study depends on the program's implementation plan. For a school-based setting with a pull-out approach, it could be that all students in a selected grade whose parents provide written consent and who themselves provide written assent will be eligible to participate in the study. In a community-based setting such as a housing authority, it could be that all youth in a certain age group residing in the housing complex will be eligible.[17]

Once written parent consent and youth assent for the study is obtained, youth will be asked to complete the baseline survey. After the baseline survey is complete, the Abt evaluation team will randomly assign individual youth and communicate the result to the program staff, who will notify the individual prior to the start of the first *¡Cuidate!* session.

---

[15] To the extent possible, the baseline will be completed before teachers/schools know the results of random assignment. Parents and students will never know the assignment status before taking the baseline.

[16] *RtR* will be delivered by outside educators, but teachers will still need to plan for loss of class time.

[17] Note that, although the program is designed for Latino youth, in general school-based programs may not exclude youth on the basis of ethnicity. In each of the *¡Cuidate!* replication sites, program staff are targeting schools with majority Latino populations but expect to serve small numbers of non-Latino youth. Even in non-school-based settings, political considerations dictate that ethnicity not be used as an eligibility criterion.

Training, Technical Assistance and Monitoring of Random Assignment

Random assignment adds a layer of operational complexity to the implementation of any program. To help sites implement and maintain random assignment, a designated Abt Site Liaison for each site will be responsible for monitoring all aspects of random assignment.

To minimize the burden of random assignment and protect the integrity of the experiment, the Site Liaisons will provide training and ongoing technical assistance to staff involved with random assignment. In addition, we will monitor compliance with random assignment to identify problems early so they can be corrected when possible and, when necessary, properly addressed at the analysis stage (e.g., through statistical corrections for noncompliance).

Training. After grantees sign an MOU with Abt, the Site Liaison and a second team member will provide hands-on training to *SSI* program staff in using the Participant Tracking System (PTS) and facilitating random assignment. Because, in all *RtR* and *¡Cuidate!* replication sites, Abt technical staff will be responsible for implementing random assignment, the manuals and training for program, school and agency staff in those sites will stress the importance of maintaining the integrity of random assignment, and clearly explain the process by which classes and individual youth will be assigned. In all sites, agency/program staff will be trained on recruitment/consent procedures and the protection of human subjects.

Ongoing technical assistance. We will implement a robust ongoing monitoring and technical support effort after the initial training is complete. Throughout the random assignment phase, and continuing until the randomization has been completed in each site, extensive communication and monitoring will occur to ensure the integrity of the random assignment process. This will involve both site-specific assistance based on the regular monitoring of randomization and enrollment, and cross-site technical assistance through peer-exchanges and materials.

Key features of our communication and monitoring strategy include:

- A toll-free Solutions Desk (staffed by Abt Site Liaisons)
- A consistent primary point of contact on the evaluation team for each site
- Monthly check-in phone calls with sites (bi-weekly during data collection periods in school-based locations)
- Monthly site-level reports of sample accumulation

Monitoring and documenting the results of random assignment. The evaluation team will monitor compliance with random assignment to ensure that it is carried out as planned, and that local sites are adhering to the results of random assignment. Monitoring also provides a way to document any instances of crossover or non-participation ("no-shows") so that adjustments can be made during the analysis phase.

For *SSI*, we will regularly track electronically the progress of randomization and services, including:

- How is the sample accumulating? (i.e., the number of new sample members added each month and their treatment/control status)
- Is everyone in the program data system "accounted for," i.e., in a treatment group, in a control group, or excluded from the study for some mutually agreed upon reason?

- Is contact and demographic information (that was to be collected in the PTS) complete?
- Are control group members receiving *SSI* intervention services? No one randomized to the control group should have ongoing data suggesting that they have participated in *SSI*.

Monitoring random assignment in school-based sites will proceed quite differently from the process described for *SSI* sites, which conduct individualized enrollment on a rolling basis. For *RtR* and *¡Cuidate!*, Site Liaisons will review class/session rosters provided by the school or program staff at least twice during the implementation of the program: once very early in the implementation, and once toward the end of implementation. Early cross-checking can identify problems to be brought to the school's attention to prevent future crossovers. Later checking will allow us to document crossover after it occurs, and account for it in the analysis. Specific monitoring issues include:

- Are control group members receiving the appropriate classes? Rosters collected periodically should indicate that control group members are not in a class or session that receives *RtR/¡Cuidate!*
- Are treatment group members still enrolled in the class/group that receives *RtR/¡Cuidate!*? If rosters indicate that any treatment group members are not receiving the *RtR/¡Cuidate!* curriculum, this will be documented so that appropriate adjustments can be made in the exploratory impact analysis estimates of the treatment effect on the treated.

Measures for the Impact Study

A set of core measures has been developed for use across all the Federal evaluations associated with this initiative, including a baseline survey measure, two follow-up surveys, and an outline for collecting program participation data. Core measures were created so that a common metric could be used across the field, as to date there is little consistency in measures across studies[18]. These measures will be used to assess the mediators, intermediate and longer-term outcomes shown earlier, in the theory of action. Baseline, short-term, and longer-term follow-up surveys will be translated into web-based measures with audio and will be available in both English and Spanish.[19]

The use of web-based ACASI will help to ensure the privacy and confidentiality of the data. This strategy offers the capacity to capture and store data in real time; each response to a question (as it is entered) is sent immediately to a central and secure database and information is not stored on any local computer. As each survey question appears on a computer screen, an audio version of that question can be heard through headphones. English and Spanish language audio versions of each survey will be available, upon request by the respondent. As words are heard, they are highlighted simultaneously on the computer screen. The audio files may be muted if the study subject desires to just read the questions. This will help to address potential literacy issues and reinforces the notion that no one else will be around when the respondent sees or hears the survey questions to which they are responding.

---

[18] The core measures were developed by Mathematica Policy Research, Inc. with input from Abt Associates as well as the interagency working group on teen pregnancy within HHS.

[19] Though every effort will be made for sample members to complete the baseline and follow-up surveys via web, paper-pencil hard copies will also be available for all approved instruments in both English and Spanish. These will be used during baseline data collection when absolutely necessary.

Data Collection for the Impact Study

To assess the impacts of the interventions, youth in all the evaluation sites will be surveyed three times: at baseline, before the intervention begins; 6-12 months after the baseline survey (short-term impacts) and 12- 24 months after the baseline survey (longer-term impacts) (Exhibit 7 summarizes the data collection schedule for each program model).[20] To the greatest extent possible, baseline data and subsequent follow-up data will be collected using web-based ACASI technology. To maximize response rates and attachment to the study over time, survey respondents will receive an OMB-approved incentive at each survey point.

Exhibit 7: Impact Data Collection Timing and Strategy

Program Model	Length of Program Implementation	First Follow-up	Second Follow-up	Baseline Data Collection strategy	Follow-up Data Collection Strategy
Reducing the Risk	8-16 weeks	12 months after baseline	24 months after baseline	Web-based Audio Computer-Assisted with paper and pencil backup	Web-based Audio Computer-Assisted with telephone follow-up
¡Cuídate!	6 weeks	6 months after baseline	18 months after baseline	Web-based Audio Computer-Assisted with paper and pencil backup	Web-based Audio Computer-Assisted with telephone follow-up
Safer Sex	6 months	9 months after baseline	18 months after baseline	Web-based Audio Computer-Assisted with paper and pencil backup	Web-based Audio Computer-Assisted with telephone follow-up

Baseline data collection

In *SSI* replication sites, trained clinic staff will obtain youth consent and, where indicated (i.e., when parents accompany a minor to the clinic) parental consent. In school-based replication sites, school staff will assist in obtaining active parental consent and student assent to participate in the evaluation. Parental consent will be obtained at the beginning of the study for possible participation in the program and for the baseline and all subsequent data collections. We will not re-consent parents at any subsequent time. Youth, on the other hand, will be asked to assent at baseline and to re-assent before completing each of the two subsequent surveys.

In school-based settings (*Reducing the Risk* and *¡Cuídate!*), we will prepare a final survey roster of all youth at each school for whom we have received parental consent and student assent, and who are expected to complete the baseline questionnaire. We will work with schools to determine dates and venues for conducting survey administration with "consented" youths. We anticipate that non-teacher school staff (e.g., nurses, guidance counselors, school support staff) designated by the school will assist with gathering youth for survey administration. Data collection staff will arrive at the school to oversee the survey, use the survey roster to take attendance, determine whether any youth are missing and exclude any not on the survey roster.

In situations where a sample member is absent for the group administration, an alternative time for individual administration will be scheduled. English and Spanish versions of the survey will also be available in hard copy format, for use in the event that unanticipated technical "glitches" occur at the time of administration. The hard copies are designed to look like the web version

[20] The evaluation research design, consent, assent, and data collection procedures were approved by the Abt Associates Institutional Review Board (IRB) for all sites and by additional local IRBs associated with the study sites as needed (e.g., for school or clinic-based approval).

and contain the same questions, skip and branching patterns and overall instructions as the web-based survey. Data collection staff will be trained in the procedures necessary to protect respondent privacy when paper surveys are used.

Once the respondent has completed the survey, the last screen will inform him or her "the survey is now complete". The youth will leave the computer, real-time verification of completion will be recorded in the survey database, and the youth will receive an incentive. In cases where a hard copy survey is completed, youth will place the entire questionnaire in a return envelope, seal it, and hand it to data collection staff. Staff will send the completed questionnaires to our survey contractor's office, where the questionnaires will be receipted and checked for completeness, and the data entered into the survey database.

At SSI sites, the baseline survey will be administered to eligible youth individually at the clinic, prior to random assignment. As at *RtR* sites and *¡Cuídate!* sites, the survey will be web-based and paper copies will be available in case they are necessary.

Follow-up survey data collection

For both follow-up surveys, it will be imperative to track study participants carefully and ensure that their contact information is up-to-date. The procedures will be similar across the nine sites, although we hope to have some assistance in tracking from schools in the case of school-based replications.

The follow-up data collection procedures for the clinic-based replications may differ slightly from those planned for the school-based replications, although our plan is to have the majority of surveys in all sites be web-based. Before any follow-up survey window opens, e-mail and hard copy flyers, advance letters, and evites will be sent to each participant, providing the link to the study website.

Although it may be possible for study participants in the school-based replication sites to complete the survey in school (either in the library or computer lab, or on computers brought to the school by data collection staff) at a pre-arranged time, others will need to access the survey on a home or community (e.g., local library) computer. In the case of the clinic-based replications, it will always be up to the individual respondent to identify a location in which she can complete the survey in privacy.

We anticipate that about 25% of study participants will complete the survey, without any additional contact. The remaining sample will receive reminders through telephone calls and social media, and we expect 35%-40% of participants to respond to these reminders. For the remainder, we anticipate that it will be necessary to use on-site data collectors, who will locate youth participants and encourage them to complete the survey. Data collection staff will be equipped with internet-accessible laptops and headphones and will, if necessary, meet with youth in a location of their choosing so that they can complete the survey.

We expect a 90 percent response rate to the baseline survey because survey administration will occur shortly after active parental consent is received (or, in the case of the clinic patients recruited to the study, at the time they are recruited for the study). This timing will ensure our contact data are current (no location problems) and that surveys can be administered to most youth in the location where the program takes place (for example, the school). In addition, obtaining the site's buy-in and assistance will be very important to maximizing the response rate; we will therefore invest significant effort in gaining their cooperation, minimizing burden on

sites, integrating an effective consent process, and assuring privacy to the youth participants. Grantee and school staff will be given detailed information about the surveys, how they will be administered and on what schedule, what involvement and time will be required of school and agency staff, and how data will be used and protected. Bringing sites into the process while minimizing burden will assure site support for the data collection effort.

We expect to achieve an 80 percent response rate at the second and final follow-up point (and an 86 percent or higher response rate on the intermediate follow-up survey). Eligibility for each data collection point does not require participation in the prior data collection point as long as youth assent is obtained for the current data point.

Analytic Approach

Two of the key research questions that the Teen Health Empowerment Study seeks to answer concern program impacts:

1. What are the program impacts on teen pregnancies/births, sexually transmitted diseases (STDs), and/or sexual activity (e.g., sexual initiation, contraceptive use, number of partners, etc.)?

    a. What are program impacts on intermediate outcomes such as knowledge of and attitudes towards sexual risk behavior, motivation to avoid risk behavior and negotiation skills?

    b. Do impacts differ for certain subgroups (i.e., gender, age, ethnicity, sexual experience at baseline)?

2. Do program impacts differ across the sites implementing a particular program model?

Program impacts will be estimated and reported separately for each program model. For each program model, there will be two impact reports: (1) an interim, or short-term report that will examine program impacts 6-12 months after study enrollment; and (2) a final long-term report on the program impacts 18-24 months after study enrollment. A comprehensive implementation study will provide information about the contexts in which evidence-based programs are implemented, the challenges faced in implementing them, and the aspects of program implementation that are associated with program impacts.

Because each program model represents a distinct strategy for achieving the goal of pregnancy and STI reduction, impact estimates will not be pooled across the three program models. Rather, impact estimates will be pooled across replication sites within a program model. OAH's requirements to define, measure, and adhere to fidelity to the program model mean that each replication is implementing the same core program elements. Within a program model, the random assignment and data collection procedures were also the same across all sites. The consistency of these design elements ensure that impact estimates pooled at the program level represent rigorous tests of a well-defined and consistently implemented program model.

At the end of five years, the study will provide credible evidence about long-term program effectiveness for three program models that are being widely replicated. Because there are three replications within each program model, the impact estimates will be more generalizable – going beyond a specific location, program sponsor, target population, or any other idiosyncratic aspects

of an individual implementation. Pooling data from all three replications of a program model will allow us to estimate program impacts on pregnancy with appropriate statistical power; it will ensure that the study is adequately powered to detect other behavioral impacts even if an individual site doesn't reach its recruitment target; and estimates will represent more diversity in program sponsorship, characteristics of the target population, and program settings.

To estimate program-level impacts, the team will use a regression framework to compare the average outcomes of treatment and control group members in each of the three sites implementing the program model. Pooled estimates will be calculated using weights that are inversely proportional to the variance of the site-level impact estimate. In addition, the team will test for differences in impact across the sites; if statistically significant differences are found, impact estimates will also be reported separately for each of the three sites implementing the program model.

The analysis team will conduct a "confirmatory" impact analysis to determine whether any or all of the three program models had impacts on teen pregnancies and sexual behavior. This analysis will span the interim and final reports, and will incorporate outcomes from both the short- and long-term follow up surveys. The confirmatory analysis will seek convincing evidence that each of the three program models have improved participants' behavioral outcomes past the end of the program. Confirmatory analysis uses a high standard of evidence for deciding if an intervention has had its intended effect, in order for its findings to be considered conclusive rather than merely suggestive. In particular, it is designed to avoid the statistical problem induced by testing multiple hypotheses at the same time, often referred to as the "multiple comparisons" problem. By contrast, secondary and exploratory analyses look for *suggestive* evidence of the programs' impacts on subgroups of interest and in other areas. Findings from these latter analyses, viewed as the best available evidence on potential program effects in secondary areas, can help inform policy but should not be taken as definitive. The distinction between secondary and exploratory hypotheses is that secondary hypotheses are specified in advance and include analyses for which there is strong theoretical justification, based on the program logic model, which can be conducted within the experimental design (i.e., by comparing treatment and control groups). These hypotheses will be limited in number. All other hypotheses will be categorized as part of the exploratory analysis.

In the sections that follow, we explain the overall analytic strategy, including the distinction between the confirmatory, secondary, and exploratory analysis approaches (Exhibit 8). Subsequent sections discuss the specific analytic methods to be used and other aspects of the research design.

Exhibit 8: Analytic Strategy Overview

Research Interest	Domain(s)	Measures	Multiple Comparisons Correction
Confirmatory Analysis			
Pooled impact on short-term behavior	Recent sexual behavior at the short-term follow up	Abstinence (no sex in last 90 days) Unprotected sex (engaged in unprotected sex in last 90 days)	MC correction across the two outcomes
Pooled impact on long-term behavior and pregnancy	Recent sexual behavior at the long-term follow up Pregnancy	Abstinence (no sex in last 90 days) Unprotected sex (engaged in unprotected sex in last 90 days) Pregnancy (since baseline)	MC correction across outcomes within each domain. No correction across domains. Interpretation according to guidelines in analysis report appendix
Secondary Analysis			
Site-specific behavioral outcomes	Behavior	Abstinence Risky Behaviors	Limit analysis to a small number of tests; no formal correction applied. Subgroup and site-specific impacts presented only if statistically significant differences across subgroups/sites.
Differential effects across subgroups	Behavior	Abstinence Risky Behaviors	
Intermediate Outcomes (Knowledge, Skills, Attitudes, and Intentions)	Knowledge, Skills, Attitudes, Intentions	Knowledge: composite or single outcomes Skills: composite or single outcome Attitudes: composite or single outcome Intentions: composite or single outcome	
Exploratory Analysis			
All other research interests	All	All	No formal correction applied

Confirmatory Analysis

The confirmatory analysis is designed to provide conclusive evidence on whether any or all of the three program models were effective at achieving the overall goal of changing sexual risk behavior and thereby reducing teen pregnancy. The confirmatory analysis will test hypotheses in three outcome domains—one at the short-term follow up (6, 9, or 12 months after baseline, depending on the program), and two at the long-term follow up (18 or 24 months after baseline). The short-term confirmatory findings on sexual behavior will be presented in the interim report. The long-term findings on sexual behavior and pregnancy, together with a discussion of the results of the confirmatory analysis as a whole, will be presented in the Final Report.

Each of the confirmatory domains, or sets of similar constructs, is defined in Exhibit 8. The key outcome measures within these domains, shown in column 3 of Exhibit 8, are defined as follows:

1. **Abstinence.** Encouraging youth to refrain from sexual activity entirely is one important channel through which programs might reduce the rate of teen pregnancy and STIs. This domain captures the effect on participants who are induced to *remain* abstinent (i.e. delay sexual initiation) or to *become* abstinent (for those who are sexually active at baseline). The measure of this outcome is a binary variable indicating whether the youth has engaged in sexual activity in the prior 90 days, which captures both the delay in sexual initiation for those abstinent at baseline and a return to abstinence for those who were sexually active at baseline.[21]

2. **Unprotected Sex.** The second behavioral channel through which programs could ultimately affect the rate of teen pregnancy or STIs is to reduce the rate of sexual risk taking for youth who are sexually active. The measure of this outcome is a binary variable indicating whether the youth has engaged in any sexual activity without a condom in the prior 90 days.

3. **Pregnancy.** All three program models are ultimately designed to reduce the rate of teen pregnancy, which is the final outcome in the logic model. Pregnancy is measured for each youth as a binary (yes/no) outcome indicating whether the youth has ever been pregnant (or for boys, gotten a partner pregnant).

Pregnancy is a cumulative outcome measured over the duration of the follow-up period, and will only be assessed in the final report. The behavioral outcomes in the confirmatory analysis are measured over a 90 day recall period; i.e., the 90 days preceding the survey. In order to minimize the concern that our confirmatory analysis would miss a behavioral impact that occurred early in the follow-up period but nonetheless affected pregnancy, we treat recent sexual behavior at the short-term follow up as distinct from sexual behavior at the long-term follow up. Confirmatory findings about short-term sexual behavior will be presented in the interim report. These findings will subsequently be used to help interpret the long-term findings in the final report, as we specify in the analysis report appendix.

All three domains will be tested in the confirmatory analysis without regard to the findings in any of the other domains. Formal multiple comparisons corrections will be applied within each domain (i.e., across abstinence and sexual risk outcomes), but not across domains. We do not adjust for multiple comparisons across domains because no individual finding would be interpreted as an "overall success."

Our plans for interpreting overall (short- and long-term) findings from the confirmatory analysis are described in the appendix titled "Guidelines for Confirmatory Analysis." Those plans specify that interpretation of the confirmatory findings is dependent on the pattern of results from all three confirmatory outcome domains. Therefore, although findings on short-term behavior

[21] Sexual activity is defined as sexual intercourse, oral sex, or anal sex. In 4 of the 9 sites, participants were not asked about anal sex.

will be reported in the interim report and can be interpreted as convincing evidence of each program's impact on short-term sexual behavior, we will advise readers that a final determination on overall program success should wait until short- and long-term findings are viewed and interpreted as a whole.

Secondary Analysis

The secondary analysis will involve tests of a small number of additional hypotheses that are well-supported by theory (the program logic model), are supported by the experimental study design, and are specified in advance of the analysis. These include tests for the following:

1. **Differential effects across sites.** For the two behavioral outcomes indicated in Exhibit 8, we will test whether impacts differ among sites (within program model). We will report site-level impact estimates if and only if we detect statistically significant differences in impacts across replications.
2. **Differential effects across subgroups.** For the two behavioral outcomes indicated in Exhibit 8, we will test the impact of each program model for subgroups defined by baseline characteristics. For all subgroup analyses, we will test for differences in impacts across subgroups. Findings for each subgroup will only be presented if we detect significant cross-subgroup differences in impact. Subgroups will include baseline sexual activity (ever had sex), race/ethnicity (White, Black, Hispanic, other), age (e.g. under 15 vs 15 and older), and gender (for the *RtR* and *¡Cuidate!* program models only, since *Safer Sex* is intended for females only).
3. **Intermediate outcomes.** The program models' common theory of change specifies intermediate outcomes in the domains of knowledge, attitudes, intentions and skills. We will test either a single outcome or a composite outcome (determined using subject-matter knowledge and/or factor analysis) in each of these three domains for each program model, using pooled data.

We will not make formal adjustments for multiple comparisons across secondary hypotheses in the main reporting of results. Instead, we rely on the careful pre-specification of conditions (as outlined above) and appeal to the theory of change to constrain the risk of false positives. In addition, when reporting results from the secondary analyses, we will caution the reader that secondary results have no formal controls for multiple comparisons. Finally, although we will report test results with unadjusted p-values, we will specify for the reader the number of tests that were conducted (within and across domains) and the number of false rejections that would be expected given the number of tests if there were no impact of treatment. With this in mind, we will suggest that secondary results should be interpreted as informative, but should not be interpreted as conclusive evidence of program effectiveness.

Exploratory Analysis

The exploratory analysis will encompass all other outcomes and research interests, e.g., site-level impacts or subgroup impacts on intermediate outcomes; pooled, site-level and subgroup impacts on other outcomes such as drug and alcohol use; impacts on outcomes that the study is not powered to detect such as STDs; the impact of dosage; and tests of the links in the logic model. This component of the analysis will also explore the extent to which the findings in this study

mirror those in the evidence review. We will not make formal adjustments for multiple comparisons when reporting on statistical significance. However, as in the secondary analysis, we will specify for the reader the number of tests that were conducted (within and across domains) and the number of false rejections that would be expected given the number of tests if there were no impact of treatment. The language used to describe exploratory results will be weaker than the language used for secondary results, and will typically be limited to noting that the results should be used to inform future research.

Analytic Methods

The impact analysis will examine the extent to which the TPP interventions affected each outcome. In testing for these effects, we will use two-tailed hypothesis test procedures, since we do not want to rule out the possibility that a program model might adversely affect one or more of the outcomes. Our basic strategy for estimating program impacts is to compare the outcomes of treatment and control group members. To control for any random variation across the sample in baseline measures, we will estimate regression models. Control variables will increase statistical precision (i.e., reduce the standard errors) of the impact estimates for a given sample size (Orr, 1999), reduce the sample size requirements of the study for a given Minimum Detectable Effect size, and reduce attrition bias from missing data (see Puma et al., 2009). The regression models shown below will be used to analyze all outcome variables.

As noted above, pooled impacts will be estimated separately for each program model, because the three programs differ in their strategies for delivering service and the duration and intensity of the services provided.

In all three replications of RtR, classrooms are randomly assigned within random assignment blocks to treatment or control conditions. The random assignment blocks comprise groups of classes within schools within sites that are similar to one another in the time of year that they are offered and the ages and grades of students in the classes. For these replications, we will estimate a regression model that accounts for the clustering of students within classrooms, which increases the standard errors of the impact estimates. To account for this form of clustering, we will use a multi-level modeling approach. This requires estimation of a model with the basic structure of equations 1-3 below. In this model, individual outcomes are modeled at level 1, while level 2 represents the unit of random assignment (or "cluster"). The level-1 model includes individual-level demographics and baseline measures as covariates, while the level-2 model includes a treatment indicator and dummy variables to represent the randomization blocks. Information about sites is contained within the block dummies. There are no specific model terms for sites because the block dummies are linear combinations of the site indicators. The model produces an estimate of the treatment effect that is a precision weighted average of the treatment effects within each of the randomization blocks.

(1) Level 1: $Y_{ijs} = \beta_{0\mathrm{js}} + \sum_{k=1}^{K} \beta_{kjs} X_{kij} + \varepsilon_{ijs}$

(2) Level 2: $\beta_{0js} = \gamma_{00\mathrm{s}} + \gamma_{01} T_{js} + \sum_{m=1}^{M} \gamma_{0ms} D_{mj} + \mu_{0js}$

(3) Level 2: $\beta_{\mathrm{k}js} = \gamma_{\mathrm{k}0\mathrm{s}}$

In these equations:

- Y_{ijs} is the outcome of interest (e.g. consistent condom use) for the i^{th} student in the j^{th} class, m^{th} randomization block, and s^{th} replication site;
- T_{js} is a dummy variable equal to 1 if class j was assigned to the treatment group and 0 otherwise;
- X_{kij} is the k^{th} baseline characteristic or covariate for individual i;.
- D_{mj} is a dummy variable equal to 1 if class j was randomly assigned within the m^{th} randomization block and 0 otherwise.

The coefficient γ_{01} is interpreted as the average pooled impact of the program on the outcome. Additionally, β and γ are coefficients to be estimated and ε_{ijs} and μ_{0js} are random terms. The regression covariates, X_{kij}, reflect the influence of background characteristics on the control group mean. All regression models include the following baseline covariates: age, race/ethnicity (Black, White, Hispanic, other), smoking, alcohol use, and marijuana use, grade and baseline sexual activity (ever sexually active). When available, we include the baseline measure of the outcome of interest as a covariate. We will analyze binary outcomes using linear regression models with heteroskedasticity-robust standard errors but, as a sensitivity check for confirmatory outcomes that are binary, we will also report the results from nonlinear models in appendices. Regression models of the form specified in equations 1-3 will be estimated using SAS PROC MIXED.

For the ¡Cuidate! and Safer Sex program models, in which individual sample members are randomized within randomization blocks to treatment or control conditions, we will estimate a model with the basic structure of equation 4.

(4) Level 1: $Y_{is} = \beta_{0s} + \beta_1 T_{is} + \sum_{k=2}^{K+1} \beta_{ks} X_{kis} + \sum_{m=1}^{M} \gamma_{0ms} D_{mj} + \varepsilon_{is}$

In this model:

- Y_{is} is the outcome of interest (e.g. consistent condom use) for the i^{th} individual in the m^{th} randomization block and s^{th} replication;
- T_{is} is a dummy variable equal to 1 if individual i was assigned to the treatment group and 0 otherwise;

and ε_{is} is a random term. Again, the X_{kis} represent baseline characteristics (specified above, except that for Safer Sex, gender will not be used as a covariate because all participants are female) and D_{mj} are dummy variables representing randomization blocks. In this model, β_1 represents the average pooled impact of the program on the outcome. Again, we plan to analyze binary outcomes using linear regression models, but as a sensitivity check for confirmatory outcomes that are binary, we will also report the results from nonlinear models in appendices.

Intent to Treat and Treatment-on-the-Treated Impact Estimates

We will focus our initial analysis in the interim and final reports on the impact of access to the TPP intervention. Because of the random assignment design, the crucial difference between the treatment and control groups will be *access* to TPP services: individuals in the treatment group

will have access to program services and possibly other, potentially-similar services available in the community (e.g., clinics), while control group members will have access to only those other services in the community. In the evaluation literature, the estimate of the average impact of access is referred to as the intent-to-treat (ITT) impact parameter. It measures the impact of having the opportunity to participate in the intervention on the outcomes under consideration for the average individual given access, not the average impact on program group members who actually participate in the intervention.

However, it is possible that some treatment group members will not participate in the program for a variety of reasons (e.g., because they had scheduling conflicts or moved away); these individuals are referred to as "no-shows." Likewise, a small number of control group members may manage to participate in program services, though the random assignment protocols are designed to minimize this possiblity. These individuals are referred to as "crossovers." In such circumstances, an estimate of the average impact of the treatment on the individuals who receive program services compared with what outcomes for those individuals would have been absent participation in the program can be calculated to supplement the main results and aid in the interpretation of the ITT. This approach is known as measuring the effect of treatment on the treated (TOT). If either no-shows or crossovers are at all common (e.g., above 5 percent prevalence), then we will take this approach in addition to estimating the ITT impact for the final report. (The interim report will estimate the ITT only.)

The conventional approach to estimating the TOT effect is to rescale the overall program-control group outcome difference—i.e., the ITT estimate—to reflect just those cases that receive program services if in the program group but not if in the control group. This methodology (Angrist, Imbens, and Rubin 1996) assumes that program group members who do not participate have no impact and that control group members who do participate experience the same impact as equivalent individuals in the program group. Computationally, the TOT estimator can be computed as the ITT impact estimate divided by $1-R_N-R_P$, where R_N is the program nonparticipation rate in the program group and R_P is the program participation rate in the control group. As a result, the TOT effect is larger on than the ITT effect. In practice, we will compute the TOT estimator using two-stage least squares in order to generate correct standard errors.

Estimating Subgroup Impacts

As part of the secondary analysis, the team will estimate impacts for key subgroups of participants and test for differences between subgroups, to better understand what works for whom. One example would be an analysis of effects of Reducing the Risk on sexual risk behavior by gender; that is, the team would compare the impacts for males with the impacts for females by including subgroup indicators and treatment*subgroup interaction terms in the model and testing for significance of the interaction term. Impact estimates will be presented for individual subgroups only when there is a statistically significant difference between subgroups; e.g., the impact will only be presented for the subgroup of boys if there is a statistically significant difference in impacts between boys and girls. The evaluation team will likewise report impacts for individual grantees only if a statistically significant difference in impacts is observed between them.

Missing Data Strategy and Baseline Balance Testing

Attrition poses a threat to the internal validity of the study to the extent that there is differential attrition between the treatment and control conditions. We used monetary incentives and intensive tracking to achieve the maximum possible response rate for both treatment and control groups. We will use case deletion for the few instances of missing outcome data (Puma et al., 2009).

Dummy-variable adjustment will be used to account for missing explanatory data. In the dummy variable adjustment method, missing cases are set to a constant and "missing data dummy variables" are added to the impact analysis model. Although the academic literature questions this approach for the general case of missing data (e.g. Jones 1996; Allison, 2002), it has been shown to give unbiased impact coefficient estimates when the treatment dummy is uncorrelated with the covariate(s) that have missing data—i.e., the special case of an RCT (Jones, 1996 and Puma et al., 2009). It is recommended by Puma et al. (2009) for dealing with missing explanatory covariates based on its performance in simulations of cluster RCTs.

We will assess whether attrition has affected the comparability of the treatment and control groups in the final analytic sample. To do so, we will use the analytic sample to conduct a series of baseline balance tests on key baseline variables. These results will be reported in an appendix.

The baseline balance tests will be conducted using models of the same form as the impact models (i.e., equations 1-4) but will have baseline measures as the dependent variables and no baseline covariates on the right hand side of the equations. These models will include the treatment dummy and the dummies for randomization blocks. The coefficient on the treatment dummy from these models will represent the baseline treatment-control difference, and the p-value of that difference will be used to assess its statistical significance.

Cases for which the outcome measures are missing for all three of the confirmatory outcomes will not be considered part of the analytic sample. When a case is in the analytic sample (i.e. has a non-missing value for at least one of the confirmatory outcomes) but has a missing value on another outcome, we will remove that case from the analysis of that particular outcome.

Outliers

Because key outcomes for this study are bounded (e.g., "sexual intercourse in previous 90 days" is a binary variable), it is not necessary to identify or correct outliers as would be the case for continuous outcomes. For unbounded demographic and/or explanatory variables (e.g., number of sexual partners), we will identify outliers using subject matter knowledge when necessary. Such explanatory variables will be re-coded as categorical variables (e.g., >1 partner, >5 partners, >10 partners) to mitigate the influence of outliers on analyses.

Mediation Analysis

As part of the exploratory analysis for the final report, we will conduct an analysis of mediation which serves several related purposes, but is primarily designed to answer the following broadly-defined research question:

> How are impacts on intermediate outcomes (attitudes, skills, and knowledge) linked to impacts on final outcomes (condom use, sexual activity, teen pregnancies/births)?

Our mediation analysis is designed to provide evidence on the pathways by which the three programs achieve any realized impacts on behavioral outcomes. In particular, mediation analysis can sometimes generate evidence on (1) whether improving intermediate outcomes contributes to the change in final outcomes; (2) whether certain intermediate outcomes are sufficient proxies for the ultimate outcomes that can be used in future research; and (3) the adequacy of measures of the mediating constructs (MacKinnon, Fairchild, and Fritz 2007).

The common logic model for the three interventions suggests that the programs should change behavior indirectly through certain intermediate outcomes. The purpose of the mediation analysis is to assess whether this hypothesized pathway is correct—i.e. whether the programs affect behavior *by affecting the intermediate outcomes*. (The subsequent link between behavior and pregnancy specified in the logic model is sufficiently well established that we will not assess this link as part of the meditation analysis.[22])

In particular, this analysis will explore whether the programs' impact on behavior is mediated by the following four intermediate outcome domains: knowledge (e.g., what STIs are and how they are transmitted), attitudes (e.g., toward early sexual activity, condom use, or birth control), intentions (e.g. to delay sex, use protection if sexually active) and skills (e.g., refusal/negotiation skills). These domains are proposed because they are specified as key mediator variables in the logic model for each of the three interventions. The study team will identify either a single representative outcome in each domain or create a composite outcome for each domain.

In conducting the mediator analysis, we propose to use a standard method in which we estimate three impacts:

1. Estimate the (pooled) impacts on the behavioral outcome.
2. Estimate the (pooled) impacts on the mediator.
3. Estimate the (pooled) impacts on the behavioral outcome, adjusting for the mediator.

We will then compare the significance and magnitude of coefficients across the three equations, from which we can draw conclusions about the intervention's pathway. Mediation is established when there is a significant relationship in equations 1 and 2 (i.e. the program must have an effect on the behavioral outcome, and the program must also affect the mediator); a significant relationship is found between the mediating variable and the outcome in equation 3 (i.e. the mediating variable affects behavior when controlling for other program effects); and the absolute value of the treatment effect is smaller in equation 3 than in equation 1 (i.e. the estimated effect of the program when controlling for the mediator—the part of the intervention that does not operate through the mediator—is closer to zero than the total effect of the treatment).

An important caveat regarding this analysis is that the statistical power of mediation analysis has been found to be very low, in part because of the requirement that there be a significant impact of the program on the behavioral outcomes in order for a mediation effect to exist (MacKinnon et al. 2002, 2004). To increase the statistical power of the mediator analysis, we will pool the

[22] See for example Guttmacher, 2013. http://www.guttmacher.org/media/inthenews/2013/06/05/index.html.

results across all three program models. While pooling across program models would likely be inappropriate in the experimental impact analysis, it is appropriate in the mediator analysis as long as the effects of the mediators on behavioral outcomes are the same across the three program models, or if we are simply interested in the average effects of the mediators across different interventions. For this evaluation, it is feasible to conduct a pooled analysis across the three program models because according to their logic models, the three programs share the same mediators.[23]

The mediation analysis will account for clustering of students in classes in the *RtR* program model, where intact classes are randomly assigned. Because the logic model specifies four mediators (knowledge, attitudes, skills, and motivation), we will use a multiple-mediator model to assess mediation effects (MacKinnon, 2000).

Assessing the Impact of Dosage

Another related set of exploratory analyses in the final report will consider the role of participation in selected program activities and/or dosage, to the extent that the implementation analysis and/or theory identifies specific strategies or program features as possible candidates for further analysis. For example, in the *SSI* program model, a research question of interest is whether the initial program session is sufficient to generate the full program impact, or whether the impact is increased by attending one or more booster sessions. Answering questions about dosage involves a non-experimental analysis approach that goes beyond treatment-control group outcome comparisons and has the potential to suffer from several types of selection biases. For example, youth who attend *SSI* booster sessions may have higher levels of motivation than those who do not, which would lead them to experience better outcomes even in the absence of the sessions. On the other hand, it is possible that individuals who choose to avail themselves of additional services could have worse outcomes than those who do not. A scenario that could produce such a result in this study is the following. Suppose that youth who attend *SSI* booster sessions are youth who, on average, engage in riskier behavior and that these youth choose to attend the booster sessions because of their greater perceived need for services. In this scenario, the youth who attend more sessions may be at greater risk for worse outcomes, even with the additional treatment, relative to those who chose to not attend the booster session.

For this reason, prior to conducting analyses of dosage effects, we will take the following steps to ensure that the analysis can be successfully conducted. First, we will identify specific hypotheses about dosage for each program model (e.g. whether booster sessions are necessary to achieve full program impact). Second, we will identify the two (or more) groups to be compared in the analysis, and conduct a test of baseline equivalence in a manner similar to the baseline balance testing performed on the full randomized sample. If the groups are more than .25 standard deviations apart, we will conclude that they are not reasonably well matched, and therefore we cannot rule out the possibility of strong selection bias. We will therefore not analyze the effects of dosage on those groups. If the groups *are* well balanced on observable

[23] However, abstinence is not an intended outcome of the *SSI* Intervention, so pooled analysis for that outcome would only include *RtR* and *¡Cuidate!* replications.

characteristics, we will conduct a dosage analysis, being careful to caveat our findings with the appropriate cautions about selection bias.

Anticipated Statistical Power

The determination of sample sizes for this study was guided by the principle that the pooled analysis of each program model should be powered to detect an impact at least as small as the smallest policy-relevant impacts found in previous studies of these programs. This will allow us to conduct a powerful test of whether on average, each program model yielded positive effects that are consistent with prior studies.

In this section we demonstrate the power of the study by calculating and presenting Minimum Detectable Impacts (MDIs), which are the smallest true impacts that the study has a high probability of detecting. The smaller the MDI, the greater the statistical power of the design. Here, we present MDIs for four representative outcomes, given the projections of likely sample sizes in each program model and replication. We also explore MDDIs (minimum detectable differences in impacts) between representative subgroups.

MDIs and MDDIs are a function of several factors including the ratio of treatment to control participants, the standard deviation of the outcome being examined in the absence of the intervention, and, crucially, the sample size on which the analysis is conducted. The impact analyses will be conducted primarily for the pooled collection of replication sites in each program model, and sometimes for each replication site separately in the secondary and exploratory analyses discussed above. Exhibit 9 shows the Minimum Detectable Impacts for the pooled and replication-specific analyses for each program model.

Exhibit 9: MDIs for Pooled and Replication-Specific Analyses

	Pregnancy	STI	Sex in Past 90 Days	Generic Behavioral Outcome
Pooled (Confirmatory)				
Safer Sex	3.4 %pts (34/1000)	2.0 %pts	4.4 %pts	5.1 %pts
Reducing the Risk	1.8 %pts (18/1000)	1.2 %pts	4.1 %pts	4.3 %pts
¡Cuidate!	2.0 %pts (20/1000)	1.4 %pts	4.6 %pts	5.1 %pts
Replication-Specific (Secondary and Exploratory)				
Safer Sex	---	---	7.6 %pts	9.0 %pts
Reducing the Risk	---	---	8.6 %pts	9.0 %pts
¡Cuidate!	---	---	8.4 %pts	9.0 %pts

[Alpha = 0.05, Power = 80%. R-squared for individual random assignment = 0.30; for cluster random assignment R-squared at level 1 is 0.35 and R-squared at level 2 is 0.65, based on Add Health data. ICC = 0.025. The base rate of pregnancy is 0.045 (RtR and ¡Cuidate!) or 0.132 (Safer Sex); the base rate of sex in the past 90 days is 0.35 (RtR and ¡Cuidate!) or 0.75 (Safer Sex); and the base rate of STIs is 0.02 (RtR, ¡Cuidate!) or 0.04 (Safer Sex). The generic behavioral outcome has a control group prevalence of 0.5. The random assignment ratio is conservatively assumed as 2:1 for Safer Sex and ¡Cuidate!; and 1:1 for RtR. This table assumes baseline samples of 950 individuals in each site for each program model and a survey response rate of 80%; in Reducing the Risk we assume that 54 classrooms are randomly assigned in each site.]

There is a large body of evidence regarding the impact of previous teen pregnancy prevention programs on the behavioral outcomes of interest, and the MDIs in Exhibit 9 should be viewed in this context. A recent meta-analysis of 31 studies of teen pregnancy prevention efforts found widely varying but statistically significant impacts on risky sexual behavior ranging from a low of 6.4 percentage points to a high of 40.3 percentage points for individual studies, with an overall

pooled impact of 6 percentage points for multi-component/youth development programs, which was the most successful type.[24] We have powered the pooled analysis to detect an impact on risky sexual behavior (as represented by a generic behavioral outcome with base rate similar to unprotected sexual intercourse) of between 4.3 and 5.1 percentage points, which is smaller than the impact of 6 percentage points found in this meta-analysis. There is less evidence on the likely impact of these programs on teen pregnancy, and what evidence does exist is mixed and mostly statistically insignificant—probably because of the large sample size necessary to detect such an impact. Each of the three pooled analyses is powered to detect a decrease in pregnancy of less than 45% (for example, a decrease from 44/1000 to 24/1000 for Cuidate!), which we consider a large but plausible effect.

Exhibit 9 would seem to suggest that the evaluation has a better chance of detecting the programs' impacts on teen pregnancy and STIs because the MDIs are smaller. However, our analysis suggests the opposite—that the evaluation has a better chance of detecting impacts on sexual risk behavior outcomes than on teen pregnancy and STIs. This is because the prevalence of teen pregnancy and STIs is much lower than the prevalence of risky sexual behaviors. Below, we discuss the implications of the MDIs in Exhibit 9 on the interpretation of study findings for each of the program models.

Safer Sex. For the main outcome of pregnancy, the evaluation is powered to detect an impact for this program model of 3.4 percentage points, or 34 pregnancies per 1000 participants. We estimate the base rate of pregnancies for this high-risk group (all of whom are sexually active) as 132 per 1000,[25] meaning that if the Safer Sex program reduces the rate of pregnancies from 132 per 1000 to 98 per 1000 (which is a 26% decrease), the study would be able to detect this impact. The study is less well-powered to detect an impact on the rate of STIs, largely because the base rate of STIs in the population is very low. We estimate the base rate as 4%; the program would need to reduce the rate by half this amount, or 2 percentage points, in order for the impact to be detected by the study. The pooled study's 5.1 percentage point MDI for a behavioral outcome with 50% prevalence in the control group means that the study would be able to detect the impact of 6 percentage points found in the original *Safer Sex* evaluation, if such an impact exists.

¡Cuidate!. Because participants in *¡Cuidate!* are not all sexually active, we estimate the base rate of pregnancies for this group to be lower than for *Safer Sex* participants, at 45 per 1000. The MDI of 20 per 1000 means that the pregnancy rate would have to decrease by 44% in order to be reliably detected by the pooled study—a large but conceivable impact. As with *Safer Sex*, the pooled evaluation is less-well powered to detect an impact on STIs, with an MDI of 1.4 percentage points; this represents a 70% decrease in STIs compared with the control group rate of 2.0 percentage points. The pooled analysis is well powered to detect an impact on

[24] Scher L, Maynard R, Stagner M. Interventions intended to reduce pregnancy-related outcomes among adolescents. Campbell Systematic Reviews 2006:12 (table 5, page 25).

[25] Estimated pregnancy rates are based on an analysis of data from the Guttmacher Institute. (http://www.guttmacher.org)

behavioral outcomes, with an MDI of 5.1 percentage points for an outcome with 50% prevalence in the control group.

Reducing the Risk. As with *¡Cuidate!*, we estimate the base rate of pregnancies for *RtR* participants to be 45 per 1000 individuals. The MDI of 18 pregnancies per 1000 individuals is similar to the MDI for *¡Cuidate!*, and represents a 40% decrease in the control group pregnancy rate. The MDI for STIs of 1.2 percentage points is slightly lower than for *¡Cuidate!*, but still very high—the intervention would need to reduce the rate of STIs by 60% in order to be reliably detected by the evaluation. Again, the pooled analysis is better powered to detect an impact on behavioral outcomes, of 4.3 percentage points.

Subgroups and Difference Between Subgroups. Impacts will be estimated for subgroups in the pooled analysis of each program model. Exhibit 10 shows the MDIs for subgroups constituting 33% and 50% of the study population for each of the three program models (for example males or females). The exhibit also demonstrates the minimum detectable difference in impacts (MDDI) between two mutually-exclusive subgroups, each of which constitutes 50% of the study population.

Exhibit 10: MDIs for Subgroups and MDDIs for Differences between Subgroups (Pooled Analysis)

	Pregnancy	STI	Sex in Past 90 Days	Generic Behavioral Outcome
Safer Sex				
33% Subgroup (n=950)	6.1 %pts (61/1000)	3.5 %pts	7.6 %pts	9.0 %pts
50% Subgroup (n=1,425)	4.9 %pts (49/1000)	2.8 %pts	6.2 %pts	7.2 %pts
Difference between 50% Subgroups (MDDI)	6.1 %pts (61/1000)	3.6 %pts	7.9 %pts	9.1 %pts
Reducing the Risk				
33% Subgroup (n=1,263)	3.0 %pts (30/1000)	2.0 %pts	6.9 %pts	7.3 %pts
50% Subgroup (n=1,895)	2.5 %pts (25/1000)	1.7 %pts	5.8 %pts	6.1 %pts
Difference between 50% Subgroups (MDDI)	3.5 %pts (35/1000)	2.4 %pts	8.2 %pts	8.6 %pts
¡Cuidate!				
33% Subgroup (n=950)	3.1 %pts (35/1000)	2.4 %pts	8.1 %pts	9.0 %pts
50% Subgroup (n=1,425)	2.8 %pts (28/1000)	1.9 %pts	6.6 %pts	7.2 %pts
Difference between 50% Subgroups (MDDI)	3.6 %pts (36/1000)	2.4 %pts	8.3 %pts	9.1 %pts

[Alpha = 0.05, Power = 80%. R-squared for individual random assignment = 0.30; for cluster random assignment R-squared at level 1 is 0.35 and R-squared at level 2 is 0.65, based on Add Health data. ICC = 0.025. The base rate of pregnancy is 0.045 (RtR and ¡Cuidate!) or 0.132 (Safer Sex); the base rate of sex in the past 90 days is 0.35 (RtR and ¡Cuidate!) or 0.75 (Safer Sex); and the base rate of STIs is 0.02 (RtR, ¡Cuidate!) or 0.04 (Safer Sex). The generic behavioral outcome has a control group prevalence of 0.5. The random assignment ratio varies between 1:1 and 2:1 depending on the program model. This table assumes baseline samples of 950 individuals in each site for each program model, with a total of 54 classes randomly assigned per site in Reducing the Risk.]

These estimates indicate how large the subgroup impacts and differences in impacts would need to be for the evaluation to be 80-percent certain to flag them as statistically significant. For example, the MDDI for STIs between two 50-percent subgroups for the Safer Sex program model is 3.6 percentage points for all three replications combined. This means that, for example, if the program reduces the rate of STIs for one of two equal-sized subgroups from 4 percent to 0.4 percent and has no effect on the rate for the other subgroup, the evaluation will have an 80 percent chance of detecting that a difference in impact of this magnitude exists between the two subgroups. The ability to detect only fairly sizeable differences in impact magnitude is typical of all but the largest impact evaluations; samples would need to be very much larger to significantly increase the probability of detecting differences in impacts between subgroups.

Reporting

The analysis team will prepare two reports of findings for each of the three program models: an interim report based on analysis of the short-term follow-up survey, and a final report describing analysis of the long-term follow-up survey and presenting conclusions drawn from outcomes in both survey waves. This reporting strategy is summarized in Exhibit 11.

The *final report* will follow the confirmatory strategy outlined in this document to determine the overall effectiveness of each of the three programs based on 18-24 months of survey measurements. The confirmatory analysis includes behavioral outcomes and pregnancy measured at the long term as well as sexual behaviors measured at the short term. The final report will also include the secondary analysis, including tests for differential effects by site and subgroups of interest; and exploratory analysis encompassing other outcomes of interest. An assessment of dosage to examine whether program effects differ with levels of participation will be included in the final report, as will a mediation analysis to examine whether program impacts on behavior (if any) are mediated by intermediate outcomes. The final report will be completed in early fall 2016.

The *interim report* will describe an early assessment of program impacts after 6-12 months. It will include an analysis of program effects on key short-term behavioral outcomes, intermediate outcomes, services received, and whether early impacts differ by site or subgroup. The interim report's findings on the key short-term behavioral outcomes specified in Exhibit 8 can be considered confirmatory evidence of short-term behavioral impact. The interim report will also include findings from the implementation analysis (not described in this document), such as the amount of program services offered and average number of sessions attended by participants. This report will be completed in summer 2015.

Exhibit 11: Reporting Strategy

Reporting Wave	Analysis Components	Timeline
Final	Program impacts Differences across sites Subgroup analysis Dosage analysis Mediation analysis	Fall 2016
Interim	Early program impacts Services received Subgroup analysis Implementation	Summer 2015

In both reports, we will report impact findings in tables showing the control group mean, the regression adjusted impact estimate (and its standard error and p-value), and the difference between these two as the inferred regression-adjusted treatment group mean. An example of a table shell for the knowledge domain is shown in Exhibit 12:

Exhibit 12: Example Table Shell for Knowledge Outcomes

Knowledge About Condoms, Birth Control, and STIs				
	Level of Knowledge			
Measure	Treatment	Control	Difference	p-value
Scale items (Average proportion of correct answers; For individual items: % selecting correct answer)				
Pregnancy Risk Knowledge Scale (4 items)	XX%	YY%	ZZ%	0.pp
Unprotected sex can lead to pregnancy	xx%	yy%	zz%	0.pp
Birth control takes effect immediately	xx%	yy%	zz%	0.pp
Condoms decrease pregnancy risk	xx%	yy%	zz%	0.pp
Birth control decreases pregnancy risk	xx%	yy%	zz%	0.pp
STI Risk Knowledge Scale (12 items)	XX%	YY%	ZZ%	0.pp
HIV persists for life	xx%	yy%	zz%	0.pp
HPV vaccine is available	xx%	yy%	zz%	0.pp
STIs can be cured by taking medicine	xx%	yy%	zz%	0.pp
STIs can be transmitted by the healthy-looking	xx%	yy%	zz%	0.pp
STIs increase risk of HIV transmission	xx%	yy%	zz%	0.pp
1 in 4 teens get STI each year	xx%	yy%	zz%	0.pp
You can get an STI from oral sex	xx%	yy%	zz%	0.pp
Condoms decrease the risk of HIV	xx%	yy%	zz%	0.pp
You cannot get HIV from unprotected sex “once or twice” without a condom	xx%	yy%	zz%	0.pp
Condoms decrease the risk of gonorrhea	xx%	yy%	zz%	0.pp
Birth control pill decreases the risk of HIV	xx%	yy%	zz%	0.pp
Birth control pill decreases the risk of gonorrhea	xx%	yy%	zz%	0.pp
Sample size	n,nnn	n,nnn		

SOURCE: Abt calculations from TPP Survey.
NOTES: For individual items, column percentages represent the proportion answering either “I am sure about correct answer” or “I think correct answer.” For composite scales (pregnancy risk knowledge and STI risk knowledge), the percentage represents the average proportion of correct answers across individuals and items.

For binary outcomes (e.g., condom use), we will report impacts as percentage point differences between the treatment and control group means. For all other outcomes, we will show impact estimates in their original metric and additionally convert impact estimates to effect sizes (by dividing by the impact estimate by the control group standard deviation) and report these in a separate column.

References

Allison, P.D. (2002). *Missing Data*. Thousand Oaks, CA: Sage University Paper No. 136.

Angrist, Joshua D., Guido W. Imbens and Donald B. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association*, 91, 444-472.

Bloom, H. S. 1984. Accounting for no-shows in experimental evaluation designs. *Evaluation Review* 8(2), 225-246.

Centers for Disease Control and Prevention. 2011a. *U.S. teenage birth rate resumes decline* (NCHS Data Brief 58). Hyattsville, MD: National Center for Health Statistics.

Centers for Disease Control and Prevention. 2011b. *Winnable battles: Teen pregnancy* [Web page]. http://www.cdc.gov/WinnableBattles/TeenPregnancy/index.html

Cheah, B.C. 2009. Clustering standard errors or modeling multilevel data? Unpublished manuscript.

Denton, J., Tsai, C., & Chevrette, P. 1988. Effects on survey responses of subject, incentives, and multiple mailings. *Journal of Experimental Education* 56, 77–82.

DiCenso, A., Guyatt, G., Willan, A., & Griffith, L. (2002). Interventions to reduce unintended pregnancies among adolescents: Systematic review of randomized controlled trials. British Medical Journal, 324(7351), 1426–1430

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. 2003. A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research* 18, 237-256.

Ilkramullah, E., Barry, M., & Manlove J. 2011. Facts at a glance [Fact sheet]. http://www.childtrends.org/Files/Child_Trends-2011_04_14_FG_2011.pdf

Jones, M. (1996). Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regressions. *Journal of the American Statistical Association*, 91, 222-230.

Kirby, D. (2001). *Emerging Answers: Research findings to reduce teen pregnancy.* Washington, DC: National Campaign to Prevent Teen Pregnancy.

Kirby, D. (2007). *Emerging Answers 2007: Research Findings on Programs to Reduce Teen Pregnancy and Sexually Transmitted Diseases.* Washington, D.C.: National Campaign to Prevent Teen and Unplanned Pregnancy.

Lenhart, A., Purcell, K., Smith, A., & Zickuhr, K. 2010. *Social media and young adults, Part 2: Gadget ownership and wireless connectivity* [Web page]. http://pewinternet.org/Reports/2010/Social-Media-and-Young-Adults/Part-2.aspx?view=all

Mangione, T. 1998. Mail surveys. In L. Bickman & D. Rog (Eds.), *Handbook of applied social research methods*. Thousand Oaks, CA: Sage.

Millar, M. M., & Dillman, D. 2011. Improving response to web and mixed-mode surveys. *Public Opinion Quarterly* 75(2), 249-269.

Orr, L. L. 1999. *Social experiments: Evaluating public programs with experimental methods.* Thousand Oaks, CA: Sage Publications.

Primo, D., M. Jacobsmeier, and J. Milyo 2007. Estimating the impact of state policies and institutions with mixed-level data. *State Politics and Policy Quarterly*, Vol. 7, No. 4 (Winter 2007): pp. 446-459.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. 2009. *What to do when data are missing in group randomized controlled trials* (NCEE 2009-0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Scher, L., Maynard, R., & Stagner, M. (2006*). Interventions intended to reduce pregnancy related outcomes among adolescents.* Paper prepared for the Campbell Collaboration.

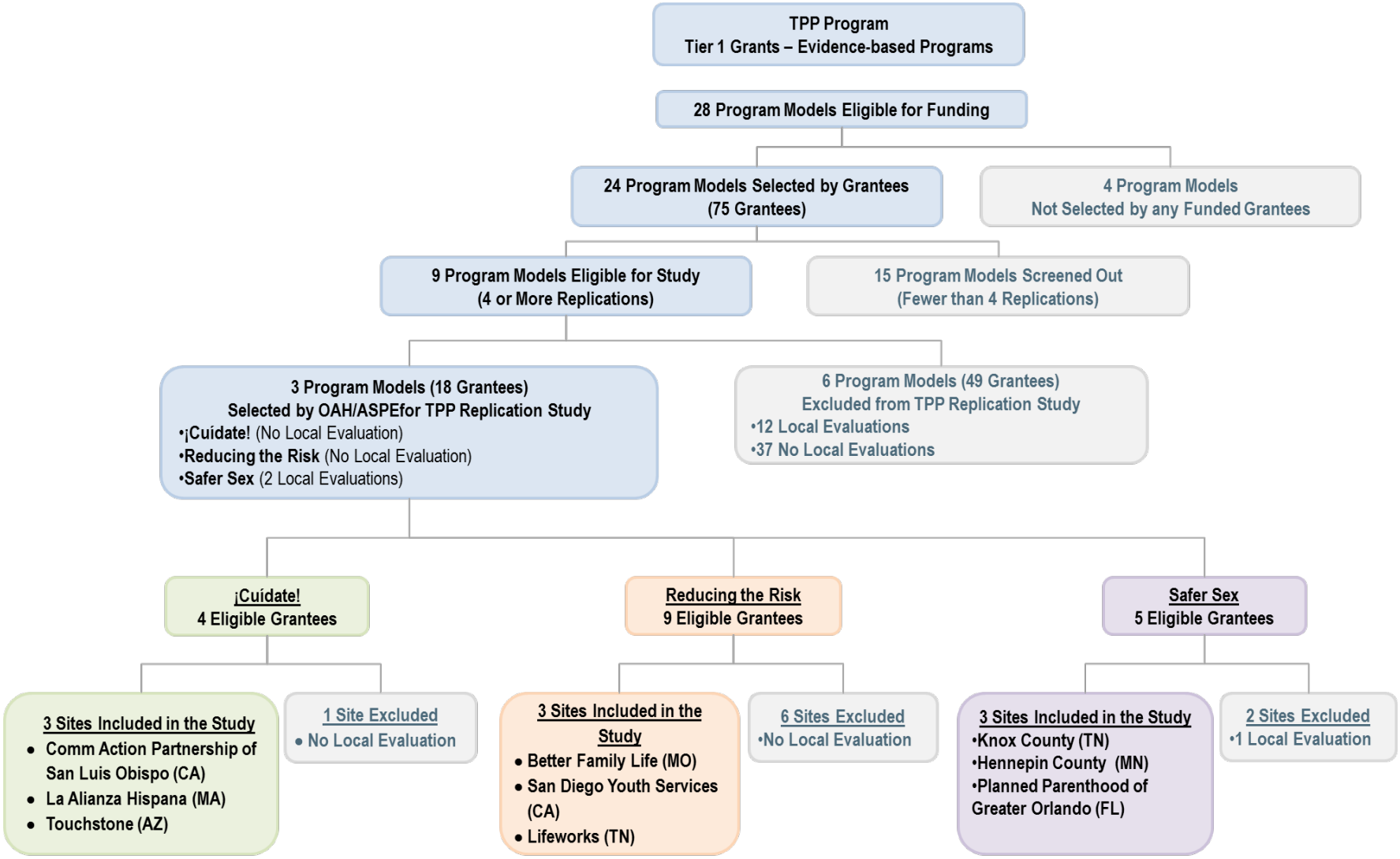
Schochet, P. Z. 2008. *The late pretest problem in randomized control trials of education interventions* (NCEE 2009-4033). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Schochet, P. Z. 2009. *Do typical rcts of education interventions have sufficient statistical power for linking impacts on teacher practice and student achievement outcomes?* (NCEE 2009-4065). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Underhill, K, Operario D, & Montgomery P. Abstinence-only programs for HIV infection prevention in high-income countries. Cochrane Database of Systematic Reviews 2007, Issue 4. Art. No.: CD005421. DOI: 10.1002/14651858.CD005421.pub2.

Weinstock, H., Berman, S. & Cates, W. (2004). *Sexually transmitted diseases among American youth: incidence and prevalence estimates*, 2000. Perspectives on Sexual and Reproductive Health, 36(1) 6-10.

Appendix A: Site Selection into the TPP Replication Study



Appendix B: Guidelines for Confirmatory Analysis in Final Report

As specified in the analysis plan, the confirmatory analysis encompasses three distinct outcome domains. The table below defines these three domains and shows which outcome measures will be included in the domain:

Domain	Definition	Outcomes
D1	**Recent sexual behavior at the short-term follow up**	**1. Abstinence in prior 90 days** **2. Unprotected sex in past 90 days**
D2	**Recent sexual behavior at the long-term follow up**	**1. Abstinence in prior 90 days** **2. Unprotected sex in past 90 days**
D3	**Pregnancy**	**1. Ever been pregnant**

Using results from each of these domains, we will interpret findings in the Final Report using the following guidelines, where "yes" means that we find a statistically significant impact and "no" means that we do not:

1. **D1 = yes, D2 = yes, and D3 = yes.** This would suggest that the intervention had an effect on sexual behavior in the both the short- and long-term, which reduced teen pregnancies.

2. **D1 = yes, D2 = yes, and D3 = no.** This would suggest that the intervention had an effect on sexual behavior throughout the follow-up period, but it would provide no evidence that the program reduced teen pregnancies over that period. There would be two possible explanations for the latter finding: First, it may be that the effect was relatively small or possibly zero--but too small to detect with a high probability. Second, maybe the effect was substantial, but the analysis, even with 80 percent power, allows a 20 percent chance of failing to detect the impact.

3. **D1 = yes, D2 = no, and D3 = no.** This would suggest that the intervention had an effect on sexual behavior in the short run but not the long run, and that the short-run effects were not enough to reduce teen pregnancy rates for the treatment group (or at least not by enough to be detected in our study).

4. **D1 = yes, D2 = no, and D3 = yes.** This would suggest that the intervention had an effect on sexual behavior in the short run but not the long run, and the short-run effects were enough to reduce teen pregnancy rates for the treatment group.

5. **D1 = no, D2 = yes, and D3 = no.** This would suggest that the intervention had no effect on sexual behavior in the short run but it had an effect in the long run—perhaps indicating that it takes time for the messages to sink in. However, the long-run effects were not enough to reduce teen pregnancy rates for the treatment group (or at least not by enough to be detected in our study).

6. **D1 = no, D2 = yes, and D3 = yes.** This would suggest that the intervention had no effect on sexual behavior in the short run but it had a positive effect in the long run—and that the long-run effects were enough to reduce teen pregnancy rates for the treatment group.

7. **D1 = no, D2 = no, and D3 = yes.** This would suggest that the intervention had no effect on sexual behavior in the short-run or the long-run, but it reduced teen pregnancy rates for the treatment group. This would be the most difficult scenario to interpret, but may indicate that there was an immediate and large short-run impact that dissipated before the observation window for the first follow-up survey, but was large enough to reduce pregnancy rates on average. It may also indicate that the program influenced one or more behaviors not captured in the survey's behavioral measures but through which the interventions could affect teen pregnancy.

8. **D1 = no, D2 = no, and D3 = no.** This would suggest that the intervention had no effect on recent sexual behavior in the short-run or long-run, and also that there is no evidence that it reduced teen pregnancy rates. This would not rule out a favorable effect on behavior in the very short run that faded out before our observation window for the first follow-up survey, but it would mean that the impact was not large enough to produce a detectable effect on pregnancy rates.