

# Identifying Programs That Impact Teen Pregnancy, Sexually Transmitted Infections, and Associated Sexual Risk Behaviors

Review Protocol | Version 5.0

## Review Author

Mathematica Policy Research  
Contact: [pprer@mathematica-mpr.com](mailto:pprer@mathematica-mpr.com)

## Review Timeline

Initial review findings: April 2010  
Updated findings: April 2012  
Updated findings: July 2014  
Updated findings: February 2015  
Updated findings: April 2016



**This page left blank for double-sided printing**

## **CONTENTS**

---

A. SEARCH FOR STUDIES .....	1
1. Call for studies .....	2
2. Keyword search of electronic databases .....	2
3. Websites of relevant research and policy organizations.....	2
4. Scan of journals and conference proceedings.....	2
B. SCREEN AND SELECT STUDIES .....	2
1. Types of participants .....	3
2. Types of programs .....	3
3. Types of research designs and data used in the analysis.....	3
4. Timeliness of the study findings.....	3
5. Types of outcomes.....	3
C. ASSESS INDIVIDUAL STUDIES .....	4
1. Study design .....	4
2. Attrition .....	5
3. Baseline equivalence .....	6
4. Reassignment .....	7
5. Confounding.....	8
D. ANALYZE EVIDENCE FOR INDIVIDUAL PROGRAMS.....	8
1. Data extraction .....	8
2. Identifying programs with evidence of effectiveness .....	9
3. Describing and summarizing the supporting research evidence .....	10
E. ASSESSMENT OF IMPLEMENTATION READINESS.....	11
F. CONFLICTS OF INTEREST .....	11
APPENDIX A. SEARCH STRATEGY .....	A-1

**This page left blank for double-sided printing**

## **TABLES**

---

1.	Summary of study quality ratings .....	4
A.1.	Keyword search databases.....	A-1
A.2.	Relevant websites .....	A-1
A.3.	Journals included in of contents search.....	A-2
A.4.	Professional associations included in scan of conference proceedings.....	A-2

## **FIGURES**

---

1.	Standard for assessing sample attrition in study quality ratings .....	6
----	--	---

**This page left blank for double-sided printing**

Since 2009, the U.S. Department of Health and Human Services has sponsored an ongoing systematic review of the research literature on programs to reduce teen pregnancy, sexually transmitted infections (STIs), and associated sexual risk behaviors. The HHS Teen Pregnancy Prevention (TPP) Evidence Review was created in response to the 2010 Consolidated Appropriations Act, which indicates that teen pregnancy prevention programs must be “proven effective through rigorous evaluation to reduce teenage pregnancy, behavioral risk factors underlying teenage pregnancy, or other associated risk factors.” Mathematica Policy Research conducts the TPP Evidence Review, which is sponsored by the Office of the Assistant Secretary for Planning and Evaluation (ASPE), the Office of Adolescent Health (OAH) within the Office of the Assistant Secretary for Health, and the Family and Youth Services Bureau (FYSB) within the Administration for Children and Families (ACF).

Mathematica updates the review findings on a periodic basis as new research emerges. Findings from the initial review of the evidence were released in spring 2010 and covered research released over a roughly 20-year period from 1989 through January 2010. The findings have since been updated about once annually to incorporate new research. Each update to the review involves the following five main steps:

1. Search for new studies released since the last update to the review
2. Screen identified studies against pre-specified eligibility criteria
3. Assess each eligible study for the quality and execution of its research design
4. Use findings from the assessed studies to identify programs with evidence of effectiveness in reducing teen pregnancy, STIs, or associated sexual risk behaviors
5. For programs showing evidence of effectiveness, assess their readiness for implementation

Each update to the review findings may include both (1) newly available evidence for programs previously reviewed and (2) evidence for new programs that prior rounds of the review did not include. When assessing newly available evidence for programs previously reviewed, the review team updates its assessment of program effectiveness by comparing the findings from the newly identified studies with the findings of those studies previously reviewed. Similarly, when assessing evidence for new programs that prior rounds of the review did not include, the review team seeks to identify and account for all currently available evidence on the program.

This document explains the specific protocol the review team follows in conducting the review. The protocol is intended in part for researchers, practitioners, and program developers wanting to learn more about the review process and how studies and programs are assessed. The protocol is also used by members of the review team as a guide for conducting each update to the review findings. The protocol has been updated over time to account for any changes in the review standards or procedures.

## **A. SEARCH FOR STUDIES**

The review team identifies new studies for each update in four ways: (1) issuing a public call for studies to solicit new and unpublished research, (2) conducting keyword searches of electronic databases, (3) reviewing the websites of relevant research and policy organizations, and (4) scanning relevant research journals and professional conference proceedings.

## **1. Call for studies**

To mark the start of each new update to the review findings, the review team issues a public call for studies through an e-mail distribution list and a posting on the TPP Evidence Review website. The call requests both (1) newly available evidence for programs previously reviewed and (2) evidence for new programs that prior rounds of the review did not include. Authors are typically given six to eight weeks to submit materials. Submissions are accepted by email.

## **2. Keyword search of electronic databases**

Additional studies are identified by conducting keyword searches of 14 electronic citation databases (see Table A.1 for a list). The searches are conducted by Mathematica's professional librarians using the following keyword combination:

*Pregnancy OR pregnant OR "HIV" OR "AIDS" OR "STD" OR "sexually transmitted"*

*OR sex\*education OR "sex education" OR abstinence*

*AND (prevention OR clinic) AND (adolescent\* OR teen\*)*

*AND (evaluation\* OR stud\*) AND (effect\* OR impact\*)*

## **3. Websites of relevant research and policy organizations**

Additional studies are identified by searching the websites of federal agencies and research or policy organizations with links to the topic of teen pregnancy prevention. The review team searches the websites of nine such agencies or organizations (see Table A.2 for list).

## **4. Scan of journals and conference proceedings**

The review team scans the tables of contents of 10 academic research journals (see Table A.3 for list) and the conference proceedings of five relevant professional associations (see Table A.4 for list). The team also searches schedules from other relevant conferences related to teen pregnancy prevention, such as the Healthy Teen Network's Conference and the National STD Prevention Conference. When potentially relevant studies or presentations are identified, the review team contacts the study author by email with information about the review and public call for studies. Authors then have the opportunity to submit their research through the public call for studies.

## **B. SCREEN AND SELECT STUDIES**

The review team screens each study identified through the literature search against a set of pre-specified eligibility criteria. These criteria account for (1) the types of participants included in the study, (2) the types of programs examined, (3) the types of research designs and data used in the study, (4) the timeliness of the study findings, and (5) the types of outcome measures examined.

## **1. Types of participants**

The review considers studies on U.S. youth ages 19 or younger. Studies with a subsample outside of this age range are considered for review if the study establishes that the majority of sample members are 19 or younger. There is no lower bound on age.

## **2. Types of programs**

The review focuses on programs that intend to reduce rates of teen pregnancy, STIs, or associated sexual risk behaviors through some combination of educational, skill-building, and/or psychosocial intervention. Programs may be delivered either one-on-one to individuals or in groups, in any type of public, private, or institutional setting. Examples include classroom-based health curricula, individualized programs delivered by health professionals in clinics or other settings, community-based or afterschool programs, and specialized programs for youth in the juvenile justice or child welfare systems. The review excludes programs that (1) focus primarily or entirely on the provision of clinical services (such as condom distribution programs) or (2) may affect sexual risk behavior and health outcomes only indirectly or through spillover effects on other outcomes (such as school dropout prevention, early childhood education, or job training programs). The review likewise excludes studies of state- or federal-policy changes, such as policies affecting access to contraception through Medicaid.

## **3. Types of research designs and data used in the analysis**

Studies must examine the effects of a program using quantitative data, statistical analysis, and hypothesis testing. The review considers both randomized controlled trials and quasi-experimental impact study designs.

## **4. Timeliness of the study findings**

To be eligible for the review, programs must have at least one impact study conducted within the last 20 years. As long as a program meets this criterion, evidence from all studies related to the program are considered for the review. However, programs for which the only impact study is more than 20 years old are excluded from the review. This “moving window” is designed to keep the review findings current and to encourage continued research on established programs.

## **5. Types of outcomes**

Studies must measure program impacts on at least one measure of sexual risk behavior or its health consequences. Measures meeting this definition fall into the following five domains: (1) sexual activity; (2) number of sexual partners; (3) contraceptive use; (4) STIs or HIV; and (5) pregnancies. Most studies use self-reported measures, but biological measures of STIs and administrative data (for example, birth records) are also considered. Measures with limitations in terms of their quality or interpretation (for example, reports from males of their female partners’ use of birth control pills or scales of behavioral risk and contraceptive use, which combine multiple measures into a single “black box” scale) are excluded from the review.

## C. ASSESS INDIVIDUAL STUDIES

Studies that meet the review eligibility criteria are assessed by teams of two trained reviewers for the quality and execution of their research designs. The first reviewer conducts a detailed assessment of the study using a modified version of the rating tool first developed by the U.S. Department of Education’s What Works Clearinghouse (WWC). The second reviewer checks and verifies the assessment for accuracy and completeness. Differences of opinion are resolved through consensus.

As a part of the assessment process, the reviewers assign each study a quality rating of high, moderate, or low according to the risk of bias in the study’s impact estimates (see Table 1). In brief, the high rating is reserved for well-implemented randomized controlled trials. The moderate rating is considered for (1) quasi-experimental comparison group designs and (2) randomized controlled trials that do not meet the criteria for the highest rating. The low quality rating is applied to studies that do not meet the review criteria for either a high or a moderate rating. The rating scheme was developed by Mathematica and approved by the U.S. Department of Health and Human Services in fall 2009.

**Table 1. Summary of study quality ratings**

Criteria category	High study rating	Moderate study rating	Low study rating
1. Study design	Random or functionally random assignment	Quasi-experimental design with a comparison group; random assignment design with high attrition or reassignment	Does not meet criteria for high or moderate rating
2. Attrition	What Works Clearinghouse standards for overall and differential attrition	No requirement	Does not meet criteria for high or moderate rating
3. Baseline equivalence	Must control for statistically significant baseline differences	Must establish baseline equivalence of research groups and control for baseline outcome measures	Does not meet criteria for high or moderate rating
4. Reassignment	Analysis must be based on original assignment to research groups	No requirement	Does not meet criteria for high or moderate rating
5. Confounding factors	Must have at least two subjects or groups in each research group and no systematic differences in data collection methods	Must have at least two subjects or groups in each research group and no systematic differences in data collection methods	Does not meet criteria for high or moderate rating

### 1. Study design

The highest study quality rating is reserved for randomized controlled trials and similar studies that randomly assigned subjects to their research groups. Studies using random assignment provide the strongest evidence that differences in the outcomes between the treatment and control groups can be attributed to the program. (Designs based on functionally

random assignment, such as alternating based on last name, date of birth, or certain digits of an identification number, are also eligible for this highest rating.)

Quasi-experimental designs with an external comparison group are eligible for at best a moderate rating. In such studies, subjects are sorted into the research groups through a process other than random assignment; therefore, even if the treatment and comparison groups are well matched based on observed characteristics, they may still differ on unmeasured characteristics. We therefore cannot rule out the possibility that the findings are attributable to unmeasured group differences. The moderate study rating is also applied to random assignment designs that do not meet other criteria for the highest rating (that is, attrition or reassignment), as explained in more detail below.

Quasi-experimental designs without an external comparison group (for example, pre-post designs) are given a low study rating. These designs are not considered for either the high or moderate rating because they offer no credible means to assess what the sample's outcomes would have been absent the intervention—a necessary condition for obtaining an unbiased impact estimate. Quasi-experimental and random assignment studies that do not meet the other criteria for a high or moderate rating are also assigned the lowest rating.

## **2. Attrition**

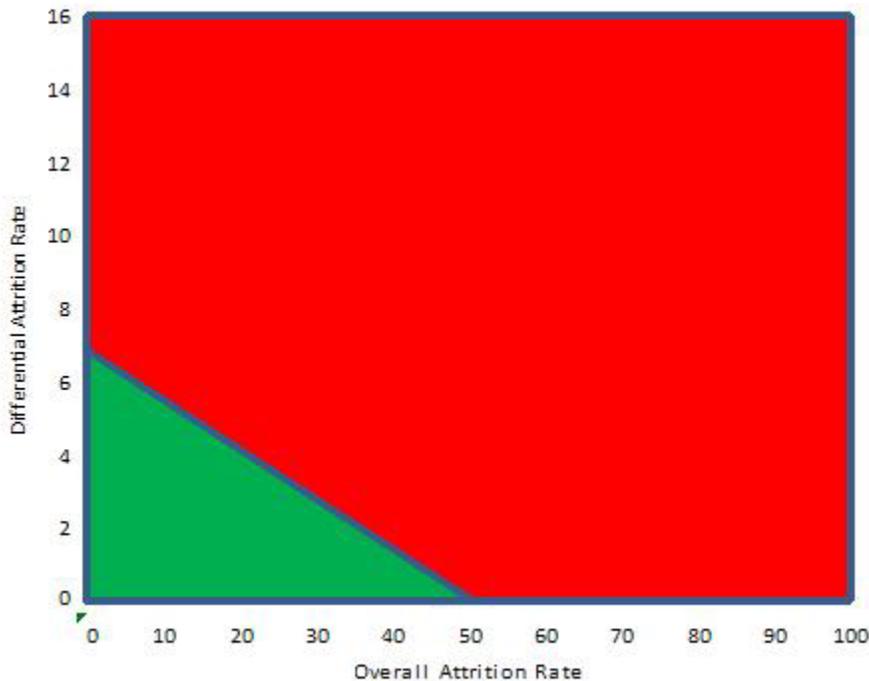
In random assignment studies, a loss of study participants can bias the study's impact estimates by creating differences in the characteristics of the treatment and control groups. Bias can arise from overall attrition (the percentage of study participants lost among the total study sample) or differential attrition (the difference in attrition rates between the treatment and control groups).

The review team assesses the level of sample attrition against standards established by the WWC. As seen in Figure 1 (next page), the WWC standards recognize a trade-off between overall and differential attrition. Namely, for an expected level of bias, studies with a relatively low level of overall attrition can meet standards with a relatively high level of differential attrition, whereas studies with a relatively high level of overall attrition require a lower level of differential attrition. Thus, the cutoff for an acceptable level of sample attrition is tied not only to the extent of overall attrition or differential attrition but rather to a combination of the two. For example, for studies with a relatively low overall attrition rate of 10 percent, the WWC standard allows a rate of differential attrition up to approximately 6 percent. However, for studies with a higher overall attrition rate of 30 percent, the WWC standard requires a lower rate of differential attrition, at approximately 4 percent. Only random assignment studies meeting the standard for acceptable combinations of overall and differential attrition are considered for the highest study rating. Random assignment studies that do not meet these standards are considered for the moderate study rating.

For cluster randomized trials, in which individuals are assigned to treatment and control conditions in groups (for example, schools or classrooms), the review team first assesses the level of attrition for the clusters or groups. If the combination of overall and differential attrition at the cluster level meets the WWC attrition standards, the review team then assesses attrition at the sub-cluster (or individual) level. Random assignment studies with low attrition at the cluster level but high attrition at the sub-cluster level are assigned the moderate study rating. Cluster

randomized trials also receive a moderate rating if sample members were added during the intervention period—for example, if a study of a multiyear pregnancy prevention program for high school students added to the sample new students who transferred into the school the year after the program began.

**Figure 1. Standard for assessing sample attrition in study quality ratings**



Source: What Works Clearinghouse. *Procedures and Standards Handbook, Version 2*. Washington, DC: U.S. Department of Education, 2008.

In calculating the rate of sample attrition, the review team compares the number of clusters and individuals at the time of random assignment to the size of the final analytic sample. Thus, any sample exclusions made after random assignment may factor into the attrition calculation. Depending on the specifics of the research design, these sample exclusions may arise from participant nonconsent, nonresponse, nonparticipation, or any number of other factors. The key determination is whether the exclusion in question presents any risk of bias to the study's impact estimates. Any sample exclusion that occurs after random assignment and presents a risk of bias will be factored into the attrition calculation.

The attrition standards are not applied to quasi-experimental studies, because the review team evaluates these studies on the basis of their final analytic samples, from which there is no attrition. This criterion is explained in greater detail below.

### 3. Baseline equivalence

In quasi-experimental comparison group studies and random assignment studies with high attrition, the use of well-matched treatment and comparison groups can minimize the risk of bias in the impact estimates. Therefore, in order to receive the moderate study rating, quasi-experimental comparison group studies and random assignment studies with high attrition are

required to demonstrate that the intervention and comparison groups were similar at baseline ( $p > .05$ , two-tailed test) on three key demographic characteristics: age or grade level, gender, and race/ethnicity. For studies with sample members at least 14 years old at baseline (or eighth grade or higher), the study authors must also establish baseline equivalence on at least one behavioral outcome measure (for example, rates of sexual initiation). This criterion is not applied to studies with younger sample members because rates of sexual risk behaviors are typically low for this age group.

Only those outcomes for which baseline equivalence is established are considered for possible evidence of program effectiveness. For example, if a study examined program impacts on three relevant outcome measures—sexual initiation, contraceptive use, and pregnancy—but established baseline equivalence for only one of the three measures (sexual initiation), the study meets the criteria for a moderate study rating, but only the impact findings for that one outcome measure (sexual initiation) are considered for possible evidence of program effectiveness. Studies are also required to control for these measures in their analyses, to ensure that any marginal differences in outcome measures at baseline did not bias the impact estimates at follow-up.

These baseline equivalence criteria are assessed on the study's final analysis sample. In some cases, studies assess equivalence for all youth who completed a baseline survey, but then present impact estimates for only a smaller subset of youth who completed a follow-up survey. These studies do not meet the baseline equivalence criteria of this review, because equivalence was not established for the smaller subset of youth on which the program impacts were based. Similarly, studies are not considered for the moderate rating if they present baseline equivalence statistics separately for subgroups defined by age, gender, or race/ethnicity, without also establishing equivalence for the full analytic sample. Some studies, for example, present baseline equivalence statistics separately for males and females or for subgroups of older and younger youth, but not for the overall combined sample.

Random assignment studies that otherwise meet the criteria for the highest rating are not required to establish baseline equivalence, because randomization is expected to produce groups that are equivalent, on average, on both observed and unobserved characteristics. Nevertheless, randomization sometimes can produce chance differences between groups and, to meet the criteria for the highest study rating, random assignment studies that show evidence of statistically significant baseline differences on behavioral outcome measures or demographics (age, race/ethnicity, or gender) are required to control for these differences in their statistical impact analyses. Random assignment studies that do not control for statistically significant baseline differences are assigned the moderate rating.

#### **4. Reassignment**

In random assignment studies, deviation from the original random assignment (for example, moving youth from the treatment to the control group) can bias the study's impact estimates. Therefore, in order for a random assignment study to meet the criteria for the highest rating, the analysis has to have been performed on the sample as originally assigned. In order to receive a high rating, subjects cannot be reassigned, based on actual treatment they received, for reasons such as contamination, noncompliance, or level of exposure. Random assignment studies that

somehow alter the original random assignment must establish baseline equivalence of their final analysis sample in order to be considered for a moderate study rating.

For similar reasons, random assignment studies cannot statistically control for measures of program dosage, participation, or any other factors that effectively alter the composition of the treatment and control groups as originally assigned. Any impact estimates resulting from such analyses are excluded from our subsequent data extraction and assessment of program effectiveness (described below).

## **5. Confounding**

In certain cases, a component of the research design or methods lines up exactly with the intervention being tested, undermining the credibility of attributing an observed effect to the intervention. For example, if a study assigns only one subject or group (for example, classroom or school) to the treatment or control condition, there is no way to distinguish the effects of the program from the particular effects of that one assigned subject or group. This can happen, for example, in quasi-experimental comparison group studies that estimate program impacts by comparing a single school or school district that implemented a pregnancy prevention program with a neighboring school or school district that did not have the program. In these cases, there is no way to distinguish the effects of the program from other characteristics of the particular school or district that implemented the program. A confounding factor can also arise from systematic differences in data collection methods for the treatment and comparison groups—for example, if program staff collect data from all subjects in the treatment group but an independent group of staff collect data from the control group. In this case, the mode of data collection cannot be separated from the effects of the intervention. Because the presence of such confounding factors severely weakens the credibility of a study's findings, a low rating is assigned to random assignment or quasi-experimental comparison group studies with either (1) only one subject or group in the treatment and control condition or (2) systematic differences in data collection procedures between the treatment and control groups.

## **D. ANALYZE EVIDENCE FOR INDIVIDUAL PROGRAMS**

All impact studies meeting the criteria for a high or moderate study quality rating are considered eligible for providing credible evidence of program impacts. For these eligible studies, the review team documents the impact estimate(s) for all relevant outcome measures, and uses this information to assess a program's evidence of effectiveness. Studies receiving a low rating are not subject to data collection and extraction, as the information provided in these studies is considered not to provide credible estimates of program impacts. The process of analyzing individual programs for evidence of effectiveness involves three sequential steps: (1) extracting information on the impact findings for each study, (2) identifying programs meeting the review criteria for evidence of effectiveness, and (3) describing and summarizing the evidence across all available studies of the program.

### **1. Data extraction**

For each relevant impact estimate from an eligible impact study, the review team collects and records the following information: the name and description of the outcome measure, length of follow-up, analytic sample used to estimate the program impact (full sample or subgroup of interest defined by (1) gender or (2) sexual experience at baseline), the reported statistical

confidence interval or associated standard error of the estimate, the reported p-value or other associated test statistic, and statistical significance level as reported by the study authors. The review team extracts this information only for eligible outcome measures as defined in the review protocol.

In the case of random assignment studies with multiple follow-up periods, this information is documented only for follow-up periods meeting the standard for low sample attrition. For follow-up periods not meeting the attrition standard, the information is treated as if it was based on a moderate quality study and documented only if the study establishes baseline equivalence for the analysis sample of that follow-up.

The review team documents all of this information as the author(s) reports it. For example, studies can report the magnitude of the impact estimates in many forms—as log-odds ratios, differences in probabilities, or effect size units—and the review team documents each magnitude as it is reported. To help users of the review make sense of these estimates and better understand the magnitude of program effects, the review team encourages study authors to report both an unstandardized and a standardized estimate of magnitude for each impact estimate, regardless of the level of statistical significance. In some cases, the review team may also follow up with study authors to request missing information on program effect sizes.

## **2. Identifying programs with evidence of effectiveness**

Based on the information collected and extracted from the eligible impact studies, the review team identifies programs meeting the review criteria for evidence of effectiveness. These criteria require a program to have at least one impact study showing evidence of a favorable, statistically significant impact on at least one outcome measure within one of the eligible outcome domains, for either the full analytic sample or a subgroup defined by (1) gender or (2) sexual experience at baseline. The eligible outcome domains are (1) sexual activity; (2) number of sexual partners; (3) contraceptive use; (4) STIs or HIV; and (5) pregnancies. In addition, the study cannot show evidence of any adverse, statistically significant impacts on any outcomes in these domains.

Statistical significance is assessed with a two-tailed hypothesis test and a specified alpha level of  $p < .05$ . For studies in which the unit of assignment is a group (or cluster) of individuals (for example, schools or classrooms), study authors must appropriately adjust statistical significance tests for the correlation in measurement among individuals within the same group (intra-cluster correlation). If the tests are not appropriately adjusted, the review team may follow up with study authors to request adjusted estimates. If adjusted estimates are unavailable, the evidence in question will be excluded from the review.

Although commonly featured in the literature, evidence from subgroups defined by sexual activity at follow-up is not considered when assessing program effectiveness. As with other endogenous subgroups that are defined by behavior emerging after the start of the program, the composition of those who are sexually active at follow-up may be affected by program participation. As a result, even with an experimental design, the treatment and comparison groups within such subgroups may lack equivalence, leading to biased estimates of a program's impact for these groups.

### **3. Describing and summarizing the supporting research evidence**

For programs meeting the review criteria for evidence of effectiveness, the review team describes and summarizes the research evidence across all available studies of the program. Some programs have been evaluated only once and so have evidence from only a single impact study. For these programs, the review team's summary of the evidence is limited to the evidence from a single study. Other programs have been evaluated in multiple, separate studies. For these programs, the review team compares and summarizes the evidence across all the available studies.

For each program, the review team describes and summarizes the evidence in each of the five eligible outcome domains: (1) sexual activity; (2) number of sexual partners; (3) contraceptive use; (4) STIs or HIV; and (5) pregnancies. For each outcome domain, the program's evidence of effectiveness is classified as falling into one of the following four categories:

1. Positive impacts: Evidence of uniformly favorable impacts across one or more outcome measures, analytic samples (full sample or subgroups), and/or studies.
2. Mixed impacts: Evidence of a mix of favorable, null, and/or adverse impacts across one or more outcome measures, analytic samples (full sample or subgroups), and/or studies.
3. Indeterminate impacts: Evidence of uniformly null impacts across one or more outcome measures, analytic samples (full sample or subgroups), and/or studies.
4. Negative impacts: Evidence of uniformly adverse impacts across one or more outcome measures, analytic samples (full sample or subgroups), and/or studies.

The review team makes these assessments separately for each of the five outcome domains. As a result, a program may be classified as having "positive impacts" in one domain but "mixed impacts" in another domain. In addition, programs are classified in these categories only for the domains on which they have been evaluated. For example, if a program has been evaluated for impacts on sexual activity but not pregnancy, the review team classifies the program's evidence of effectiveness only for the domain of sexual activity.

When comparing findings across multiple studies of the same program, the review team bases this comparison whenever possible on the estimated effect sizes and confidence intervals reported in the individual studies. In particular, when multiple studies examine program impacts on a common outcome, the review team records the effect size and associated confidence interval reported in the first study establishing the program's evidence of effectiveness. For subsequent studies, the review team then examines whether the effect sizes reported in these studies fall within the confidence interval reported in the initial study. The review team uses these effect size comparisons to assess the consistency of findings across studies, rather than the statistical significance tests and p-values reported in the individual studies. This comparison of effect sizes is possible only for studies examining common outcome measures and using comparable analytic methods. The review team does not compare effect sizes across different outcome measures or across studies that estimate impacts using different analytic methods.

## **E. ASSESSMENT OF IMPLEMENTATION READINESS**

For programs meeting the review criteria for evidence of effectiveness, the review team conducts an independent assessment of each program's readiness for implementation. This assessment is based on the team's review of available program materials and documents. The team also requests input from program developers and distributors about availability of implementation materials and resources.

On the basis of this assessment, the team calculates an implementation readiness score comprised of three component scores: (1) curriculum and materials (has defined curriculum with lesson plans and/or activities, has defined core or required components, and/or has facilitator's guide or instructions), (2) training and staff support (available formal pre-implementation training by qualified trainers and/or available supplemental training or ongoing technical support available), and (3) fidelity monitoring tools and resources (has defined logic model, defines fidelity guidelines and benchmarks, and/or offers monitoring and evaluation tools). The component scores are combined into a total score, which ranges from 0 to 8, with higher scores indicating the programs most ready to implement.

## **F. CONFLICTS OF INTEREST**

Members of the review team are not allowed to assess studies they were involved in designing or conducting. The review team does not otherwise face any potential conflicts of interest.

**This page left blank for double-sided printing**

**APPENDIX A. SEARCH STRATEGY**

---

**Table A.1. Keyword search databases**

Database
Academic Search Premier
CINAHL with Full Text
Cochrane Methodology Register
Cochrane Central Register of Controlled Trials
Cochrane Database of Systematic Reviews
Database of Abstracts of Reviews of Effect
Dissertation Abstracts
Education Research Complete
ERIC
Health Policy Reference Center
Mathematica's in-house E-journals database
MedLine
PsycInfo
SocINDEX with Full Text

---

**Table A.2. Relevant websites**

---

1	<a href="#">Advocates for Youth</a>
2	<a href="#">Centers for Disease Control and Prevention (HIV/STD Prevention Research Synthesis)</a>
3	<a href="#">Guttmacher Institute</a>
4	<a href="#">Healthy Teen Network</a>
5	<a href="#">National Abstinence Clearinghouse</a>
6	<a href="#">National Abstinence Education Association</a>
7	<a href="#">National Campaign to Prevent Teen and Unplanned Pregnancy</a>
8	<a href="#">Sociometrics (Program Archive on Sexuality, Health, and Adolescence)</a>
9	<a href="#">Child Trends (LINKS database)</a>

---

**Table A.3. Journals included in table of contents search**

---

- 1 American Journal of Maternal Child Nursing
  - 2 American Journal of Public Health
  - 3 Archives of Pediatric and Adolescent Medicine
  - 4 Journal of Adolescent Health
  - 5 Journal of AIDS Education and Prevention
  - 6 Journal of Consulting and Clinical Psychology
  - 7 Journal of School Health
  - 8 Perspectives on Sexual and Reproductive Health
  - 9 Public Health Reports
  - 10 Sexually Transmitted Diseases
- 

**Table A.4. Professional associations included in scan of conference proceedings**

---

- 1 American Public Health Association
  - 2 Association of Maternal and Child Health Programs
  - 3 Society for Prevention Research
  - 4 Society for Research on Adolescence
  - 5 Society for Research in Child Development
-