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

Review Protocol | Version 7.0

October 2023

Review Author
Mathematica
Contact: tpper@mathematica-mpr.com

Review Timeline
Initial review findings (version 1.0 of standards): April 2010
Updated findings (version 2.0 of standards): April 2012
Updated findings (version 3.0 of standards): July 2014
Updated findings (version 4.0 of standards): February 2015
Updated findings (version 5.0 of standards): June 2016
Updated findings (version 5.0 of standards): April 2018
Updated findings (version 6.0 of standards): April 2023
Pending updated findings (version 7.0 of standards): Spring 2024

This page left blank for double-sided printing.

CONTENTS

This page left blank for double-sided printing.

TABLES

FIGURES

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 conducts the TPP Evidence Review (TPPER), which is sponsored by the Office of the Assistant Secretary for Planning and Evaluation (ASPE), the Office of Population Affairs (OPA) (previously, the Office of Adolescent Health) 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 periodically. The latest round of review captured studies published through early 2022. 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, describe their components and implementation requirements

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) scanning the tables of contents of relevant research journals, and (4) reviewing citations in recently published literature reviews and meta-analyses.

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 TPPER 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. During the call for studies period, OPA and FYSB are also invited to submit studies from their recently funded evaluation grants.

2. Keyword search of electronic databases

Additional studies are identified by conducting keyword searches of 15 electronic citation databases (see Table A.1 for a list). The searches are conducted by Mathematica's professional librarians using the following keyword combination:

(((((HIV[ti] OR AIDS[ti] OR pregnan[ti] OR sexually transmit*[ti] OR STI[ti] OR STD[ti] OR birth[ti]) AND prevent*[ti]) OR (sex[ti] AND educat*[ti]) OR (sexual*[ti] AND (initiate*[ti] OR minorit*[ti] OR health[ti] OR risk avoid*[ti])) OR sexually[ti] OR "sex ed"[ti] OR abstinence[ti] OR abstain[ti] OR "risky sex" OR "teen pregnancy") AND (adolescen*[ti] OR teen*[ti] OR youth*[ti] OR student*[ti] OR minor*[ti] OR young[ti])) AND (Treatment[ti] OR interven*[ti] OR component[ti] OR trial[ti] OR trials[ti] OR program*[ti] OR evaluat*[ti] OR random*[ti] OR quasi*[ti] OR matched[ti] OR review[ti] OR systematic[ti])) NOT (Africa OR "Africa"[Mesh])*

3. Scan of journals

The review team also scans the tables of contents of 13 academic research journals (see Table A.2 for list) to identify studies that might be eligible for review.

4. Review citations in recently published literature reviews and meta-analyses

The review team also reviewed the reference lists of recently published literature reviews and meta-analyses to identify studies that were not picked up in the literature search or call for papers.

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.

In addition, this review focuses on the impact of well-defined components or combinations of components of programs that intend to reduce teen pregnancy, STIs, or associated sexual risk behaviors. To be eligible for review, a component must be (1) a clearly defined practice, procedure, policy, support, or organizational structure, potentially with documented steps for implementation with fidelity to facilitate replication; and (2) capable of being implemented independently, in conjunction with, or integrated into a TPP intervention. Examples of components that could be eligible for review include practices such as in-class condom demonstrations and text-messaging as an enhancement to a well-defined TPP program.

The review excludes programs or components 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 with follow-up data collection 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 with evidence of effectiveness 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;[1] 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 original rating scheme was developed by Mathematica and approved by the U.S. Department of Health and Human Services in fall 2009; the rating scheme was most recently revised in 2022.

[1] STI testing is an eligible outcome as long as the test is not provided as part of the intervention.

Table 1. Summary of study quality ratings

Criteria category	Features of studies with the high study rating	Features of studies with the moderate study rating
Study design	Random or functionally random assignment	Random assignment design with high attrition or reassignment; Quasi-experimental design with a comparison group
Attrition	Random assignment studies that do not exceed What Works Clearinghouse standards for overall and differential attrition (cautious assumption)	Random assignment studies that exceed What Works Clearinghouse attrition standards; Attrition is not assessed in quasi-experimental designs
Baseline equivalence	Not assessed – samples are assumed to be equivalent by virtue of random assignment and low levels of sample attrition	The equivalence of the research groups is demonstrated at baseline, and systematically adjusted for in impact analyses
Reassignment	Analysis is based on original assignment to research groups	Not assessed, given the baseline equivalence requirement described above that ensures equivalence of the research groups
Confounding factors	At least two subjects or groups in each research group and no systematic differences in data collection methods	At least two subjects or groups in each research group and no systematic differences in data collection methods

Note: Studies that do not achieve the high or moderate rating are given a "low" study 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

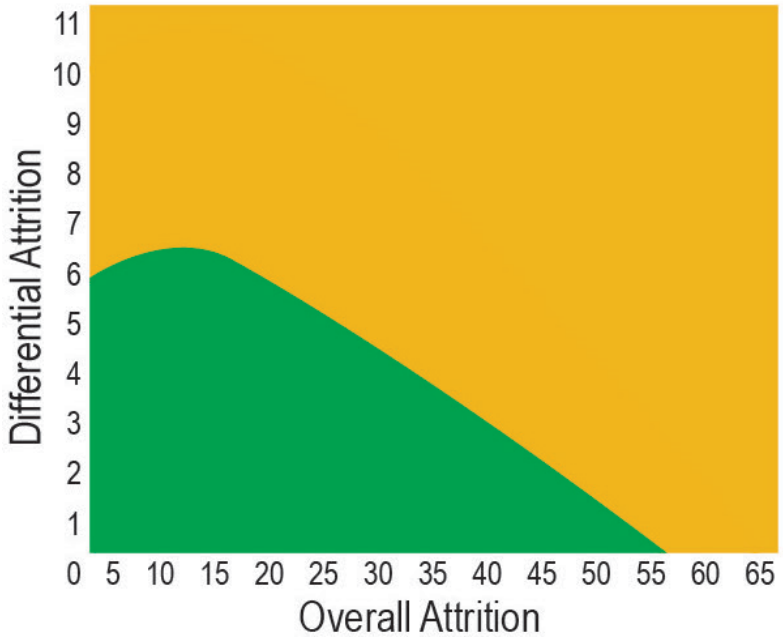
In random assignment studies, a loss of study participants can bias the study's impact estimates by creating differences in the characteristics of the treatment and control groups. Bias can arise from overall attrition (the percentage of study participants lost among the total study sample) or differential 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 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 using the cautious assumptions 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 among non-attriting clusters. Random assignment studies with low attrition at both levels are eligible for the high rating. Random assignment studies with high attrition at either level must demonstrate baseline equivalence of the analytic sample to be eligible for the moderate study rating.

In addition, cluster randomized trials that include sample members in the impact analysis who were not included in the sample at the time of random assignment (in other words, they joined the sample after random assignment) may also be required to demonstrate baseline equivalence of the analytic sample to be eligible for the moderate study rating. This requirement is enforced in contexts where the unit of assignment could potentially be exploited by joiners (for example, when classrooms within a school are the unit of assignment and a student may join a particular classroom in order to get the intervention).

Figure 1. Standard for assessing sample attrition in study quality ratings (WWC cautious attrition assumption)



Source: WWC. *Procedures and Standards Handbook, Version 4.1*. Washington, DC: U.S. Department of Education, 2020.

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[2]. 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 concerns about sample composition change (for example, studies with high attrition, reassignment, or individuals included in the analysis who may have selected/joined a cluster based on an attractive intervention), 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

[2] Sample attrition is calculated based on observed data (in other words, unimputed data) in the analytic sample, regardless of any approach used to impute or address missing data.

concerns about sample composition change are required to demonstrate that the intervention and comparison groups were similar at baseline 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.

TPPER will use the following approach to determine if samples satisfy the baseline equivalence requirement: If the reported difference of a specified baseline characteristic is greater than 0.25 standard deviations in absolute value, based on the variation of that characteristic in the pooled sample of treatment and control group members, TPPER considers the treatment and control groups to be nonequivalent.

Depending on the size of the baseline difference, TPPER may require a statistical adjustment in the analysis. TPPER has slightly different rules for statistical adjustment requirements for demographic characteristics and baseline measures of the outcomes:

- For demographic characteristics, when differences in the specified baseline characteristics are greater than 0.05 and lower or equal to 0.25 standard deviations, the analysis must include a statistical adjustment to meet the baseline equivalence requirement.[3] Differences of less than or equal to 0.05 standard deviation require no statistical adjustment.
- For baseline measures of the outcome, any difference lower than or equal to 0.25 standard deviations must be statistically adjusted for.[4]

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.

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

[3] When demographic characteristics are presented for multiple categories (for example, multiple races or genders), the assessment of baseline equivalence will be based on the modal category.

[4] Including baseline measures on the left side of the regression equation (a difference-in-differences approach) will be an allowable means of statistically adjusting for baseline differences for continuous and count outcomes, but not for dichotomous outcomes (unless the authors justify the pre-post correlation for these outcomes); the pre-post correlation for dichotomous outcomes rarely exceeds the $r=.60$ threshold typically required for a difference-in-differences adjustment to effectively adjust for baseline differences, which is why this approach is allowable only for continuous or count outcomes.

statistics separately for subgroups defined by age, gender, or race/ethnicity, without also establishing equivalence for the full analytic sample on which they estimated program impacts. 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. Finally, studies must demonstrate baseline equivalence of their analytic samples for various outcomes using unimputed baseline data. When there are multiple analytic samples, studies should ideally present baseline equivalence for each analytic sample. To avoid overburdening study authors, TPPER reviewers may assess baseline equivalence using information for a sample of individuals that differs slightly from the sample of individuals used to produce a finding, (for example, due to item-level nonresponse on a survey) provided the difference in samples falls below the threshold for high attrition.

Some impact evaluations (notably, quasi-experimental studies and random assignment studies with high levels of attrition) use various statistical techniques to equate treatment and comparison groups at baseline. These techniques include (among others) (1) estimating propensity scores and limiting the analytic sample to the subset of observations that match well on the scores or (2) calculating (entropy or inverse-propensity) weights and using those weights to produce more credible impact analyses. These equating approaches are likely to improve baseline equivalence, and thus reduce confounding, relative to comparing the original (unweighted or unmatched) treatment and control groups.

Studies using these types of equating approaches are potentially eligible to receive a moderate rating if they satisfy the following requirements:

- The equating approach must include only exogenous variables in the calculation of the score or weight used to equate groups. Exogenous covariates are variables the treatment status will not potentially affect. Although TPPER does not prescribe which variables to include in an equating approach, if a review determines that a model included potentially endogenous variables (such as level of engagement with the program), then all results based on the model will receive a low rating.

- The success of the equating approach must be assessed by comparing the effect size differences between the matched or weighted analytic sample for all required baseline variables. Per the baseline equivalence standards discussed before, if the effect size differences are greater than 0.05 and lower than or equal to 0.25 standard deviations, the analysis must include an appropriate statistical adjustment. Differences less than or equal to 0.05 standard deviations do not require a statistical adjustment (except for a baseline measure of the outcome, which does require adjustment). Differences greater than 0.25 standard deviations do not meet the baseline equivalence requirement.

- Adjusting for the propensity (or other equating) score by itself (for example, by including it as a covariate in the impact model) is not sufficient when statistical adjustments for baseline measures of the outcomes or demographic characteristics

are required. When a required covariate in a matched sample design requires statistical adjustment, the impact model should directly adjust the required covariate.

- If a study uses weighting approaches to equate groups, the study must document that the sum of the weights in the analytic sample is less than or equal to the number of observations in the analytic sample. This step is necessary to guard against artificially enhancing the precision of the standard errors and impact estimates that often result from outlier weights.

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.

6. Analysis considerations

Some studies contain multiple follow-up periods and conduct impact analyses that incorporate more than one follow-up assessment in a single analytic model (for example, in a repeated measures, difference-in-differences, or growth curve analysis). In such situations, reviewers will separately assess the potential internal validity threats associated with the evidence contributing to each follow-up assessment. That is, reviewers will conduct separate attrition assessments at each point in time included in the impact analytic approach (as needed) and assess the baseline equivalence of the analytic samples contributing to each point-in-time impact estimate (as needed). Although some studies will report differences in trends as an estimate of program effectiveness, TPPER will prioritize the point-in-time differences in outcomes as the focal effect size statistics of interest and will therefore assess internal validity threats at each point in time even in studies that do not report impacts separately for each time point. If the authors do not provide this information, TPPER will query authors for effect size information at each point-in-time.

Study authors must handle missing data appropriately, regardless of design. The most common and straightforward method researchers use when data are missing is to simply remove observations with missing data from the samples they analyze and conduct a complete-case analysis. But other methods for handling missing data are sometimes used, including imputation (replacing observations with guesses as to the most reasonable value) or maximum likelihood (creating a statistical model to account for the missing data), and these alternate approaches may provide more credible estimates of program effectiveness than complete-case analyses. The WWC Standards Handbook Version 4.1 lists five acceptable approaches to handle missing data, along with standards for how RCTs and QEDs with missing outcome or baseline data should be handled (WWC 2020). When studies present credible analyses that align with WWC's acceptable approaches for handling missing data, TPPER will allow such studies to receive a moderate or high rating, depending on other features of the study design and execution.

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[5] 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. 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.

[5] In a multi-site evaluation, site-specific impacts can be considered full sample contrasts when they are presented as intentional feature of the study design (for example, the sites represent different replication settings, the report mentions that it pre-registered a plan to report site-specific impact estimates, etc.).

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 (see Colman 2012).

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 study, the review team first describes and summarizes the findings 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 study's findings are classified as falling into one of the following six categories (Table 2). In addition, on the TPPER website, the TPPER team will report the magnitude of the observed impacts for all outcomes that meet TPPER standards for a moderate or high rating, when information is available. This will include information on both unstandardized impacts and standardized effect sizes, which will be calculated by the TPPER team when feasible.

Table 2. Domain categorization rating for individual studies

Domain rating	Criteria
Favorable findings	• Two or more favorable impacts and no unfavorable impacts, regardless of null findings
Potentially favorable findings	• At least one favorable impact and no unfavorable impacts, regardless of null findings
Indeterminate findings	• Uniformly null findings
Conflicting findings	• At least one favorable and at least one unfavorable impact, regardless of null findings
Potentially unfavorable findings	• At least one unfavorable impact and no favorable impacts, regardless of null findings
Unfavorable findings	• Two or more unfavorable impacts and no favorable impacts, regardless of null findings

To characterize the evidence for a program (which may include findings from multiple studies), the review team uses the following rating approach for each outcome domain (Table 3):

Table 3. Domain categorization rating for programs (potentially pooling across studies)

High-level rating	Specific rating	Criteria
Favorable	**Favorable evidence**: Strong evidence of favorable findings with no overriding contrary evidence	• Two or more studies show Favorable findings, AND • No studies have **Inconsistent findings, Potentially unfavorable findings,** or **Unfavorable findings**
	Potentially favorable evidence: Evidence of a favorable findings with no evidence of adverse findings	• At least one study shows **Favorable findings** or **Potentially favorable findings**, AND • No studies have **Inconsistent findings, Potentially unfavorable findings,** or **Unfavorable findings**
Null	**Indeterminate evidence**: No affirmative evidence of findings	• All of the studies show **Indeterminate findings**
Conflicting	**Conflicting evidence**: Evidence of conflicting (both favorable and unfavorable) findings	• At least one study shows **Inconsistent findings** OR – At least one study shows **Favorable findings**, or **Potentially favorable findings**, AND – At least one study shows **Unfavorable findings**, or **Potentially unfavorable findings**
Unfavorable	**Potentially unfavorable evidence**: Evidence of unfavorable findings with no overriding contrary evidence	• At least one study shows **Unfavorable findings** or **Potentially unfavorable findings**, AND • No studies have **Inconsistent findings**, **Potentially favorable findings**, or **Favorable findings**
	Unfavorable evidence: Strong evidence of unfavorable findings with no overriding contrary evidence	• Two or more studies show **Unfavorable findings**, AND • No studies have **Inconsistent findings**, **Potentially favorable findings**, or **Favorable findings**

The review team makes these study and program assessments separately for each of the five outcome domains. As a result, a program may be classified as having "favorable evidence" in one domain but "conflicting evidence" 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.

E. DESCRIBING PROGRAM COMPONENTS AND IMPLEMENTATION REQUIREMENTS

For programs meeting the review criteria for evidence of effectiveness, the review team will report information on program components and developer-intended and study-level implementation requirements. For studies of components that meet review criteria of effectiveness, the review team will report implementation information separately from the information on full programs. The review team will use information from full programs and component studies to develop implementation profiles summarizing information about each program or component that will appear on the TPPER website. That site will report data elements on broad topics:

For the program implementation profile:

- **Program overview**. This section will summarize information about the program and its intended population and setting, and developer contact information.
- **Program components**. This section will summarize information about the ingredients of the program, including its objectives and goals, content, and instructional methods.
- **Implementation requirements and guidance**. This section will summarize information about program structure and timeline, staffing requirements, fidelity guidelines, and allowable adaptations.

For the component implementation profile:

- **Component description.** This section will summarize information about the component and its intended population, setting, dosage, mode, and the larger program it was implemented as part of or supplemental to (if applicable).
- **For more information.** This section would note the component's developer or distributor, describe whether the component is publicly available or available for purchase and, if so, how to access it. The section would include information about languages available, other materials available (for instance, fidelity or adaptation guidelines), and so on.

The purposes of the implementation profiles are to 1) document the most updated implementation information available for programs or components that demonstrate evidence of effectiveness; 2) provide context to help understand study research findings; 3) help the public learn more about a particular TPP program or component and whether it might be a good fit for their community, setting, and population; and 4) begin to disaggregate core program and implementation components. The target audience for the profiles will be practitioners and/or program administrators/staff as well as researchers.

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
Science Direct
SocINDEX with Full Text

Table A.2. Journals included in table of contents search

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