Evaluation Technical Assistance Brief
for OAH Teenage Pregnancy Prevention Grantees

September 2017

Selecting Benchmark and Sensitivity Analyses

Despite best efforts to be independent and impartial, and to let the data speak clearly, researchers must make difficult decisions that play a role in the findings that they produce from their impact evaluations. After specifying a research question about the effectiveness of a program, researchers face many decisions about how to operationalize the analysis —for example, how to clean contradictory data or which statistical approach they should use to estimate the program's impact. Such decisions are challenging, because there are often several justifiable but competing approaches, each of which can lead to different results. Researchers could stumble on a potentially erroneous result that depends on an arbitrary modeling decision. As a consequence, they might inadvertently highlight a finding that does not reflect the true effect of the program, rather the finding is an artifact of their analytic decisions. Findings that are highly sensitive to research methods are considered less credible (Leamer 1985).

This brief discusses how to estimate and present a set of analyses that reveal how sensitive the results are to the researcher's analytic decisions. We propose using a benchmark analysis *that will serve as the primary answer to the research question and a set of* sensitivity analyses *that will summarize how that answer might change under different assumptions. We draw attention to common situations in which teen pregnancy prevention (TPP) researchers make decisions that could influence their findings, and we highlight how sensitivity analyses can help protect the integrity of the results and paint a more comprehensive picture about the effects of a program. This approach also helps avoid any appearance of "fishing for results" or "p-hacking," which arises when researchers privately conduct many different analyses but publicly report only the most favorable or statistically significant results (Wasserstein and Lazar 2016).*

Planning an analytic approach for estimating impacts

Developing an approach for estimating impacts first requires the researcher to define the research question of interest. Articulating the research question helps set the stage for the correct approach to answering it—that is, the correct way to analyze the data to understand the effect of a program. A typical research question for a TPP effectiveness study takes the general form: "What is the impact of intervention X on outcome Y for population Z?" That is, the research question defines the program being tested for a given outcome with a particular population. Depending on the design, the research question might add more details, such as the impact at *a particular time point after the intervention*.[1]

Researchers face many more decisions about how to operationalize the details of the analysis to answer the research questions. These decisions include choosing how to code missing or contradictory data, constructing variables and selecting an estimator for the parameter of interest. For many decisions, there is not always a correct choice, giving rise for the need to conduct multiple analyses that show how results differ under alternative but justifiable choices. The next section, representing the bulk of this brief, covers decision points that are common for TPP researchers.

One key consideration when conducting multiple analyses is how to effectively present findings given there is not an objectively best approach. By focusing on the results of a benchmark analysis as the primary analytic approach—and conducting and acknowledging the results of sensitivity analyses—researchers can emphasize the overall takeaway from the study as well as discuss the robustness of the findings. This general approach helps maintain a parsimonious presentation for general audiences, while still including the necessary details for more critical and technical readers. See the final section of the brief for more information on presentation and interpretation of results from various analyses.

> To avoid the appearance of fishing for results, we suggest specifying the plans for both the benchmark and sensitivity analyses before conducting the analyses, specifically in an impact analysis plan.

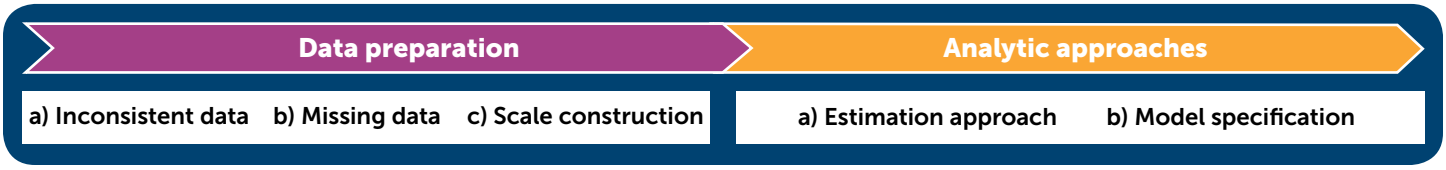
Decisions in analytic approaches

This section discusses decision points associated with the data preparation and analytic approach phases of estimating program effectiveness in an impact evaluation. Figure 1 outlines the key

[1] The research question itself pins down several modeling decisions—for example, the treatment parameter of interest. See Table A.1 in Appendix A for some examples.

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA
OFFICE OF ADOLESCENT HEALTH

Figure 1. Process for selecting benchmark analysis and sensitivity analyses



decision points that we discuss in more detail. Please note that this is not an exhaustive list; we focus primarily on issues commonly encountered in TPP evaluations.

We first explore decision points at the data preparation phase and then turn to decision points about analysis. At each phase, researchers face choices and decision points about how to proceed, each of which could lead to different results. We therefore highlight various approaches for how researchers might address a given situation. We also recommend which approaches to use as benchmark and sensitivity specifications and suggest other information to present to help justify the appropriateness of a given decision.

Inconsistent data

TPP surveys often include two or more questions that could yield inconsistent responses, particularly when administering a survey on paper. For example, a participant might state in one question that they have never had sexual intercourse but state in a later question they first had sexual intercourse at age 18. These inconsistencies can be addressed using several different approaches (See Goesling 2012 for more information). We describe three common approaches below. Table 1 shows the tradeoffs among these approaches and offers some recommendations.

1. ***Ignore inconsistency.*** In this approach, the inconsistent responses are left unmodified.
2. ***Treat inconsistent data as missing.*** In this approach, all inconsistent data are recoded to missing. Depending on the analytic approach, this might mean that the analytic sample excludes people with inconsistent data from the estimate of program effectiveness.
3. ***Logically impute subsequent responses using the first response.*** In this approach, the first response is assumed to be correct and subsequent inconsistent responses are recoded to be consistent with the first response. In the earlier example, the first

Table 1: Pros and cons of approaches to handling inconsistent data

Approach	Pros	Cons	Considerations
(1) Ignore inconsistency	• Maintains sample size • Easy to implement	• Leads to impact estimates that might contradict one another • Uses some data that must be incorrect because at least one of the inconsistent responses is wrong • Could lead to biases if the intervention affects inconsistencies	• From a face validity perspective, it might be difficult to argue that this approach is best and a viable benchmark, given that some data are incorrect. This approach might be more appropriate as a sensitivity analysis.
(2) Treat inconsistent data as missing	• Produces impact estimates that are internally consistent with one another • Requires no assumptions about which response is correct	• Reduces sample size • Could lead to attrition bias, especially when prevalence of inconsistency differs between treatment groups	• The Youth Risk Behavior Survey uses this approach
(3) Logically impute subsequent responses using the first response	• Maintains sample size • Impact estimates internally consistent with one another • Mimics a computerized administration with skip patterns	• Assumes that the first response is correct • Could lead to biases if the intervention affects inconsistencies	• Most compelling if combining data from paper and pencil administration with data from computerized administration that has built-in skip patterns

Overall recommendation

There is no one best approach for handling inconsistent survey responses, as any approach requires the researcher to make assumptions and assertions about why data are inconsistent. At the minimum, the researcher should acknowledge the issue of inconsistent responses and apply a systematic approach and clearly defined decision rule to all inconsistent data in the study. They should justify the appropriate-ness of the rule and offer an approach that has face validity to a critical reviewer. For example, in justifying an approach, the researcher should provide data about the prevalence of inconsistent responses by treatment condition and by survey mode. For studies with a high prevalence of inconsistent data, we recommend treating inconsistent data as missing as the benchmark approach and conducting a sensi-tivity analysis with at least one of the other methods.

response was that the participant did not have sexual intercourse. To make the response to the question about age at first intercourse consistent with the first response, age at first intercourse would be recoded to missing or not applicable. This approach mimics computerized assessments in which the skip patterns are programmed—for example, if the respondent reported they had not had sexual intercourse, the computerized assessment would automatically skip the question about their age at first intercourse.

Missing data

Nearly all impact evaluations have some missing data, so researchers must plan a strategy for addressing this common problem. Missing data can lead to at least two types of bias: (1) it can produce biased impact estimates if respondents in the treatment and control groups differ from each other in systematic ways, and (2) it can produce bias in the generalizability of the observed impact estimate if the respondents contributing to the analytic sample differ from the target population of interest.

We briefly highlight three common approaches for preparing and analyzing data from a TPP evaluation with missing data below. Deke and Puma (2013) provide a more thorough discussion of these issues and guidance on ways to describe the extent and nature of missing data as well as analytic approaches for addressing missing outcome or baseline data. Table 2 describes pros and cons of the different approaches and recommends an approach for TPP impact studies. In addition, Appendix B provides more detail on information required to meet evidence standards in situations when either baseline or outcome data (or both) are missing.

1) ***Complete case analysis without imputation.*** All observations with missing data for any item (either baseline or outcome) included in the analysis are excluded. This approach is also known as listwise deletion.

2) ***Complete case analysis after imputing missing baseline data to a constant value.*** Under this approach, first missing baseline data are imputed to a constant, and a dichotomous variable is created indicating which observations had the missing variable imputed to a constant value. In the impact regression analysis, include both the baseline variable (with complete data) and the dummy indicator as covariates. Observations with missing outcome data are typically excluded from the ultimate impact regression analyses.

Table 2: Pros and cons of approaches to handling missing data

Approach	Pros	Cons	Considerations
(1) Complete case analysis without imputation	• Simplest approach • Sample used for demonstration of baseline equivalence aligns 1:1 with sample used to show impacts	• Potentially the smallest analytic sample size, which might attenuate statistical precision • Only provides an unbiased estimate of program effectiveness if data are missing completely at random	• This approach might be advisable as a method to meet U.S. Department of Health and Human Services evidence standards if there are extensive missing data (see recommendation below).
(2) Complete case analysis after imputing missing baseline data to a constant value	• Potentially improves precision relative to (1) • Is unbiased under less severe missing data processes—if missing baseline data are related to other observed variables (that is, the data are missing at random), then this approach will be unbiased, an improvement relative to (1)	• Sample used to demonstrate baseline equivalence is either a smaller sample than the analytic sample (that is, not a 1:1 correspondence), or uses imputed data for demonstration. Both of these limitations may call into question the equivalence of the analytic sample by an evidence review.	• This approach might be advisable if there is not extensive missing data on key baseline variables; if a substantial proportion of the sample is missing key baseline variables, an alternate approach might be preferable.
(3) Multiple imputation or maximum likelihood imputation	• Maintains sample size • Impact estimates internally consistent with one another • Mimics a computerized administration with skip patterns	• More computationally complex to implement • Might be difficult to communicate to non-research audience	• This approach might be advisable if there is not extensive missing data on key baseline variables; if a substantial proportion of the sample is missing key baseline variables, an alternate approach might be preferable.

Overall recommendation

Overall, we recommend using either Approach 2 or 3 for the benchmark analysis because these approaches enable a larger sample size (potentially increasing statistical precision) and control some types of biases that might exist under Approach 1. See Deke and Puma (2013) for tradeoffs between these approaches and the most appropriate contexts for their use. That said, we strongly recommend conducting Approach 1 (complete cases analysis without imputation) as a sensitivity analysis, because some evidence reviews might prioritize a complete case analysis as the most credible way to mitigate composition bias concerns

3) ***Multiple imputation or maximum likelihood imputation.*** Under these approaches, all missing values (baseline and outcomes) are imputed using a model, and standard errors in the ultimate impact analysis are adjusted to reflect the uncertainty of the imputation.

Please note that different approaches may be useful depending on the study design and levels of sample attrition observed. For example, if the study is a randomized controlled trial (RCT) with high levels of sample attrition at the unit of assignment, or is a quasi-experimental design (QED), then it will only be eligible for the moderate evidence rating. Approach 1 can only meet evidence standards with a moderate rating if the complete case sample is shown to be equivalent at baseline on required variables of interest—it might be necessary to trim the analysis and create a matched sample for this complete case analysis to be eligible to meet U.S. Department of Health and Human Services evidence standards. See Cole and Agodini (2014) for more information on this topic.

Scale construction: Analyzing multiple items that capture the same construct

TPP researchers often use multiple items that capture a single underlying construct such as knowledge, attitudes, intentions, or self-esteem. For example, Rosenberg's (1965) self-esteem scale requires respondents to report whether they strongly agree, agree, disagree, or strongly disagree with ten statements, each of which represents a different facet of self-esteem, for example, "On the whole, I am satisfied with myself," or "I feel that I have a good number of qualities."

There are two general approaches to estimating impacts with multiple, related items. One approach is to analyze each item as a separate outcome. A second approach is to combine the information from multiple items into a scale and then estimate a single impact for that composite scale. This second approach can improve reliability and statistical power as well as reduce the number of separate impact estimates, potentially making it easier to report findings and limiting the necessity of multiple comparisons adjustments. The second approach is valid when the items adequately capture the same underlying construct both empirically and theoretically (in that the items naturally fit together and can be described or labeled as one construct).

There are several alternative methods for analyzing a set of related measures, suggesting the necessity of sensitivity analyses. We describe four commonly used methods below. Table 3 outlines some tradeoffs among these methods and offers recommendations.

1. ***Analyze items separately.*** In this approach, each item is treated as a separate outcome. An impact analysis could compare the average value of each item between the treatment and control group. Because some items, such as the Rosenberg scale (a Likert scale with response categories from 1 to 4 representing strongly disagree to strongly agree categories), do not have natural units, researchers sometimes analyze items as the fraction of people who report a particular answer category, for example, an item from the Rosenberg scale could be reported in terms of the fraction of people who strongly agree with each statement. Please note: For TPP evaluations, scaled measures are not eligible to meet U.S. Department of Health and Human Services (HHS) standards. Behavioral outcomes should be presented as clearly interpretable measures, for example sexual initiation or number of sexual partners (Mathematica Policy Research, 2016).

2. ***Create a simple scale by averaging or summing across items.*** If a group of items captures the same construct, one approach is to create a simple average of the items. In the example of the self-esteem scale, in which each item takes a value of 1 to 4, one possible scale is the average or sum of these values across items. When using this approach, researchers should report a measure of reliability (for example, Cronbach's alpha) that gives a sense of how related the items are to each other, because if the items are not related, then it is not valid to combine them in this way. See Nunnally and Bernstein (1994) for a comprehensive discussion of scale creation and reliability.

3. ***Create factor scores.*** Factor scores provide another way to combine items into a single variable. This approach proceeds in two steps. First, the researcher estimates a factor model that provides an estimate of how each item depends on the underlying construct (Gorsuch 1983). Second, the researcher uses the model to calculate a score for each individual, which is an estimate of the value of the latent construct. The score is a weighted average of the individual items such that items with less measurement error receive higher weights.

4. ***Estimate a structural equation model of the construct and intervention impact.*** Structural equation modeling is another approach for using multiple items to measure a construct. This method is similar to using factor scores because it also uses a factor structure to define the constructs of interest. The difference is that the model can simultaneously estimate how the items relate to the construct and the impacts of the intervention on the construct. The models typically make assumptions about the distributions of the factors. See Hoyle (2012) for a comprehensive guide to structural equation modeling.

Approaches to adjusting for covariates

Researchers must also choose whether to adjust for baseline covariates, that is, whether to include baseline covariates in the outcome equation when estimating the impact of the intervention. Assuming a valid experiment, adjusting for covariates is not strictly necessary because experimental randomization balances the observed and unobserved characteristics between the treatment and control groups. Nevertheless, adjusting for covariates has two potential benefits: (1) it can account for chance imbalances in observable characteristics between the treatment and control groups, potentially giving a more reliable estimate; and (2) it can improve statistical power

Table 3: Pros and cons to various approaches for handling constructs

Approach	Pros	Cons	Considerations
(1) Analyze items separately	• Easy to implement and explain • Enables each item to be impacted differently by the intervention	• Might be harder to present if there are many items • Could require multiple hypothesis corrections if there are many items within a single domain	• For TPP, "black box" scales are not eligible to meet HHS evidence standards.
(2) Create a simple scale by averaging or summing across items	• If all items relate to the same latent construct, improves precision/reliability (that is, reduces measurement error) relative to presenting items separately • Reduces the number of multiple hypothesis corrections conducted across individual items • Can concisely summarize findings • Can make the impacts more comparable with other research if using an existing scale	• All items within a construct enter the composite measure with equal weight, which might not be optimal to maximize reliability of the construct • Could obscure effects on individual items	• When items are combined into a scale/construct, it is important to accurately label the construct based on the content of the items that comprise it. • With this approach, we recommend also presenting measures of reliability, such as Cronbach's alpha, especially in cases where the scale has not been previously validated.
(3) Create factor scores	• Same pros as (2) • Enables more informative items to enter the scale with a higher weight • Improves precision relative to presenting items separately and using a simple scale • Provides information about how the underlying construct is related to each item of interest	• Might be less transparent than (1) or (2) • Some technical readers might see this approach as data driven, rather than theory driven • Could obscure effects on individual items	• With this approach, we recommend presenting measures of reliability, validity, and model fit. • Might have to start with exploratory factor analysis as first step and then confirm factor structure using theory-driven confirmatory factor analysis to convince readers of technical merit of approach.
(4) Estimate a structural equation model of the con-struct and interven-tion impact	• Same pros as (3) • Better corrects for measurement error than scales or factor scores	• Typically imposes a functional form assumption on the data generating process • Might not fit with the impact model estimation, for example, if a linear probability model is preferred for dichotomous outcomes • Could obscure effects on individual items • Most computationally intensive	• Same considerations as (3)

Overall recommendation

The most appropriate method for presenting impacts on constructs depends on the audience and the extent to which items capture the same underlying construct. The most basic method—presenting the items separately—is arguably the most transparent. The other methods all increasingly reduce measurement error in the underlying construct but also come at a cost of increasing complexity in estimation and possibly transparency. In general, using a composite (for example, an average across items or factor score) is most appropriate when the items are shown to have a high reliability within the study (for example, when they are highly correlated with each other, which produces a high Cronbach's alpha) or are part of an existing scale that has been validated. Using a composite also reduces the number of outcomes, which might ease explication and reduce the necessity of multiple hypothesis corrections.

by explaining variation in the outcomes. Covariates that are highly correlated with the outcome variables will increase statistical power the most (Schochet 2008). Table 4 outlines some of the tradeoffs between different approaches and offers some recommendations.

Approaches to account for blocked designs

In blocked designs, randomization occurs separately within mutually exclusive subsamples, or blocks. This approach is analogous to conducting mini-experiments within each block. For example, blocks could be schools where individual students are randomly assigned to condition separately within each school. Or in a clustered design, schools could be randomly assigned to condition within districts, where districts serve as blocks. In some designs, blocks correspond to baseline demographic variables. For example, the blocks could be based on gender so that randomization occurs separately for males and females. In this case, the blocking ensures that the treatment and control groups have the same number of males and females. We describe several ways that the empirical analysis can account for blocked designs below. Table 5 outlines some of the tradeoffs and offers some recommendations.

Table 4: Pros and cons of approaches to covariate adjustments

Approach	Pros	Cons	Considerations
(1) No adjustments	• Easy to implement and explain • Produces a valid treatment effect (assuming no imbalance)	• Potentially susceptible to chance imbalances • Reduced precision if there are available baseline measures are correlated with outcomes	• If there are imbalances in the sample on key covariates (such as demographics or baseline measures of the outcome of interest), the HHS evidence review will require a statistical adjustment for these imbalances for an RCT to be eligible for the highest evidence rating.
(2) Include a small, parsimonious set of pre-specified baseline covariates	• Likely increases statistical power • Corrects for chance imbalances between treatment and control groups	• May not adjust for other variables that could influence outcomes, which would affect bias and precision of impact estimate	• Given the consideration described above, this approach might be optimal as a benchmark in the context of the HHS evidence review when the covariate set includes gender, age, race and ethnicity, and a baseline measure of the outcome of interest.
(3) Include a comprehensive set of pre-specified baseline covariates	• Same as (2) • Corrects for additional chance imbalances	• Potentially susceptible to the appearance of data fishing if comprehensive covariate list not specified in advance	• Variables to be adjusted for should have some rationale or literature to support their inclusion as theoretically or empirically related to the outcome of interest. • Variables included to offset baseline differences should be included based on a pre-specified criteria (for example, a variable being significantly different across groups at baseline).
(4) Use a pre-specified approach to iden-tify baseline covari-ates (for example, adjust for all base-line variables that are significantly different from each other at baseline)	• Same as (2) • Allows the observed data to determine the variables that will be adjusted for, rather than pre-specifying a list	• If the process incorporates nuances in the observed outcome data, rather than baseline data, then this process might produce a biased impact estimate and standard error.	• The "process" for selecting baseline variables must be pre-specified, in order for this approach to be credible. If the process uses outcome data (for example, adjusting for variables that maximize the model R^2, then this process can be criticized as a version of data mining. • The variables examined at baseline to inform the process should be those that are theoretically linked to the outcome (likely, a version of the comprehensive variable list shown in (3) above).

Overall recommendation

For any study, we recommend specifying the covariates (if any) that will be adjusted for during the impact analysis planning phase, or the process that will be used to identify the covariates that will be adjusted for. Pre-specification helps to avoid any appearance of data fishing during the analysis and reporting stage of an evaluation. We recommend picking one covariate specification as a benchmark and one other specification as a sensitivity analysis.

One suggested approach for the benchmark analysis, with the goal of meeting HHS evidence standards, is to adjust for a baseline measure of the outcome variable, gender, race and ethnicity, and age. Sensitivity analyses could adjust for imbalanced covariates or no covariates, or using a pre-specified process to determine the covariate set.

1. ***Ignore blocks (that is, treat the experiment as though it were a non-blocked design).*** In this approach, the model does not account for the blocked design. Ignoring the blocks will typically lead to a more conservative analysis (that is, inflated standard errors).

2. ***Include block fixed effects.*** In this approach, the outcome equation includes dummy variables for each block as part of the set of covariates. This approach is consistent with finite-population inference in which the estimates do not generalize to a greater sample of the blocking variable (see Appendix Table A.1 for a description). For example, if the blocks were schools, then the impacts generalize to the specific schools in the study.

3. ***Include block random effects.*** In this approach, the model includes random effects for each block. This approach is consistent with an inference in which the estimates generalize to a greater sample of the blocking variable. This approach might be appropriate if the blocks are schools, and the study team randomly sampled study schools from a larger population of schools (for example, a school district). Relative to the fixed effects approach, the standard errors will tend to be greater because they incorporate this additional source of variability in the model.

Ways to account for clustered data

The empirical analysis of clustered designs requires special consideration. In a clustered design, random assignment occurs for groups of individuals, or clusters. For example, it might involve randomly assigning entire schools to a treatment or a control group so that all students within a particular school receive the same treatment condition. Unless the analysis accounts for the clustered nature of randomization, the standard errors will be biased, and the *p*-values for tests of statistical significance will be incorrect. The bias increases with the correlation between outcomes within clusters, often measured by the intracluster correlation coefficient. For a discussion of these issues in the TPP context, see Deke (2013).

There are two common ways to account for clustering: (1) hierarchical linear modeling (HLM) and (2) robust cluster standard

Table 5: Pros and cons of approaches to accounting for blocks

Approach	Pros	Cons	Considerations
(1) Ignore blocks	• Most straightforward to explain	• Does not account for the randomization design	• This approach can be considered conservative in that the standard error will be larger than if blocks were included as fixed effects.
(2) Include block fixed effects	• Likely increases statistical power • Aligns analytic approach with the design of the experiment	• In situations with many blocks relative to the number of observations, including block fixed effects can reduce the degrees of freedom, yielding less reliably estimated standard errors—in some cases, this might adversely affect study precision relative to ignoring blocks.	• Most effective when blocks are predictive of outcomes and the number of blocks is small compared with the sample size • Sometimes blocks are defined by baseline covariates, so might not be necessary to control for both
(3) Include block random effects	• Enables the model to be representative of a greater population defined by the blocking variable	• Imposes assumptions about the distribution of blocks • Will reduce statistical precision relative to block fixed effects	• Appropriate when inference is intended for broader generalization and assuming sampling of blocks from a larger population

Overall recommendation

As with baseline covariates, we recommend specifying the approach to account for blocking during the impact analysis planning phase to avoid any appearance of data fishing, since the selection of an approach will influence the precision of the impact estimate. For most TPP studies, we recommend either ignoring blocks or including block fixed effects as the benchmark approach, and using the alternative as a sensitivity analysis. We suggest this approach because few TPP studies use random selection to enable a generalizable impact, potentially obviating the need for an analysis that includes block random effects.

The decision of which approach to lead with as the benchmark analysis can be informed by ex-ante power calculations to determine the extent to which including the block fixed effects is likely to help—please see Deke (2016) and Imbens (2011) for more info. On one hand, including block fixed effects tends to yield more precise estimates of the *treatment effect*. On the other hand, including block fixed effects will reduce the degrees of freedom used for hypothesis testing because the estimates of the *standard errors* are less reliable. All else equal, this reduction in degrees of freedom reduces the power of statistical tests. In general, including block fixed effects will help when (1) the sample size is large relative to the number of blocks, and (2) the blocks are predictive of outcomes. In cases when the blocks are defined using baseline covariates, then it might make sense to control for either the blocks or the covariates because they will be highly collinear, and therefore, including both will unnecessarily sacrifice degrees of freedom.

errors (RCSE).[2] Although both these approaches enable correlations between individuals' outcomes within clusters, they differ in some key ways. We describe these approaches below. Table 6 outlines some tradeoffs between the two approaches and some overall recommendations.

1. ***HLM.*** HLM allows for multiple levels of data—for example, a simple clustered design where clusters are randomly assigned to condition would have two levels: one for the cluster and one for the individual. Each level has an associated error term so that the model includes a cluster-level error and an individual-level error, typically assumed to be normally distributed. The cluster-level error component is the same for all individuals within the cluster, whereas the individual-level term varies within a cluster. The cluster-level term enables outcomes to be correlated within the same cluster. In the econometrics literature, this approach is sometimes called a random effects model. Raudenbush et al. (2002) provides a comprehensive treatment of HLM.

2. ***RCSE.*** RCSE is a method of calculating standard errors for ordinary least squares (OLS) regressions. In a standard OLS regression, the error terms are assumed to be independent across individuals. In contrast, with RCSE, individual error terms are correlated between individuals that are in the same cluster but are independent between individuals in different clusters. Importantly, this correlation can differ for each pair of individuals within a cluster, which contrasts with HLM approaches that assume the correlation among individuals within clusters is the same across all clusters. See Cameron and Miller (2015) for a practitioner's guide to using RCSE.

Modeling binary outcomes

Most TPP evaluations will examine research questions focusing on outcomes with binary responses, such as whether a respondent has ever engaged in sex. We recommend reporting the treatment effects for binary outcomes in terms of the marginal effect—the impact of treatment on the probability that the binary outcome equals one. The marginal effect provides a clear interpretation. For example, if the marginal effect were -0.10 in the earlier example, the interpretation is "The intervention led to a 10 percentage point decrease in the sexual initiation rate among study participants." There are two general approaches to modeling binary outcome variables to obtain the marginal effect (see Table 7):

1. ***Linear probability model (LPM).*** The first approach is to use OLS or HLM (as appropriate) to estimate an LPM, in which the treatment and other covariates have a linear effect on the probability that the variable equals one.

Table 6: Pros and cons of approaches to accounting for clustered data

Approach	Pros	Cons	Considerations
(1) HLM	• Is most efficient if the model is correctly specified • The standard errors are less biased relative to RCSE with few clusters. • Allows for blocks to be included as random effects	• Makes distributional assumptions about the error term (for example, that it is normally distributed), though these can be relaxed through non-linear approaches for hierarchical modeling (such as Hierarchical generalized linear models, or HGLM) • Does not allow variances in the error to differ between the treatment and control group	• Some reviewers might balk at the use of the linear probability model for dichotomous outcomes (described below) in the HLM context, due to the distributional assumptions about the error term. It might be useful to supplement an HLM linear probability model analysis in a clustered design with a hierarchical estimation routine that allows for non-linear relationships (HGLM).
(2) OLS with RCSE	• Does not make distributional assumptions about the error term • Allows for general forms of heteroscedasticity	• The standard errors are biased when there are few clusters.	• Some corrections and bootstrapping approaches perform well with few clusters.

Overall recommendation

The most appropriate way to account for clustering depends on the number of clusters and the desire for flexibility in estimation - specifically, a desire to understand the variability in parameter estimates across clusters. In general, the HLM approach makes more assumptions and some evidence suggests it has better small sample properties. OLS with RCSE is less restrictive in terms of assumptions, but can yield biased standard errors when there are few clusters. See Cameron and Miller (2015) for a discussion of RCSE with few clusters and rules of thumb for the number of clusters required.

Unlike some of the other decision points, the decision between HLM and RCSE is less likely to be viewed with skepticism about data fishing, and thus, it might be feasible to simply select one approach and justify its selection, rather than conducting extensive sensitivity analyses.

[2] A less common approach is to aggregate the subcluster-level data to the cluster level, and analyze the impacts on cluster-level averages. We ignore this approach because of its infrequent use in the TPP field.

2. ***Logit or probit models.*** In logit or probit models, the covariates have a linear impact on a continuous latent variable, not the binary outcome itself. If the latent variable exceeds a threshold, then the outcome equals one. These methods make assumptions about the latent variable.

Table 7: Pros and cons of approaches to modeling binary outcomes

Approach	Pros	Cons	Considerations
(1) LPM	• Most straightforward to use because the estimated coefficient directly represents the mean marginal effect of treatment • Has been shown to perform well in experimental designs (Deke 2014)	• It is possible for predicted values of the outcome to fall outside the 0-1 range, though this issue is not typically a concern for analysis of experimental designs.	• Researchers might consider accounting for the fact that the LPM will be heteroscedastic (the variance of the error term will depend on the covariates) by using an adjustment such as heteroscedasticity-robust standard errors (White 1980).
(2) Logit or probit models	• Predicted values of outcomes will fall within the 0-1 range.	• The estimated coefficient is not directly interpretable because it is not equal to the marginal effect (which depends on the covariates and differs across individuals).	• Because the marginal effect depends on the value of the covariates, we recommend reporting the percentage point impact on the predicted probability across individuals, which can be done readily with some software packages, such as Stata.

Overall recommendation

We recommend using an LPM for the benchmark analysis because it directly yields an estimate of the marginal effect and performs well in experimental designs. We suggest using either a logit or probit to conduct a sensitivity analysis, if desired. If calculating the marginal effects for the logit or probit model proves challenging, the sensitivity analysis could compare the sign and significance of the estimated coefficient with the sign and significance from the LPM.

Alternative to traditional model-based methods for estimating impacts: Design-based approaches

So far in this brief, we have focused on traditional model-based methods, such as HLM or OLS with robust cluster standard errors. These methods contrast with design-based methods, which are based on a recently developed theory for estimating impacts for randomized controlled trials (RCTs) (Imbens and Rubin 2015; Schochet, 2016).

Compared with model-based methods, the design-based methods make fewer assumptions and more explicitly incorporate aspects of experimental design into the estimators. For example, in model-based methods, the source of uncertainty in the data (that is, the error term or disturbance) is assumed to come from a particular distribution or take a particular form (for example, the error is often assumed to be normally distributed). However, in design-based methods, the uncertainty in the model comes from the randomization itself, and the properties of the error term are derived from the known features of randomization. Therefore, design-based approaches could provide a more nuanced and appropriate framework for estimating impacts when certain model-based assumptions are violated.

Design-based approaches can accommodate many of the same aspects of traditional model-based approaches, including clustering, stratifying, and adjusting for baseline differences. One potential challenge is that design-based methods might be less known by applied practitioners. That said, a recent study found that design-based estimators yield similar results and conclusions as traditional model-based methods for nine past RCTs (Kautz et al. 2017).

The design-based methods can be implemented easily with a publicly available software package called *RCT-YES*, which was developed by methodological experts with funding from the Institute of Education Sciences. *RCT-YES* is a free tool that researchers can use, in combination with standard statistical packages, to present impact findings for evaluations of interventions. The software and associated documentation are available at https://www.rct-yes. com/. Along with estimating the impacts, *RCT-YES* produces publication-quality tables and figures. The analyses conducted in *RCT-YES* and accompanying result tables and figures can provide the information required by the HHS evidence standards review team to assess the internal validity of a study (Scher and Cole 2017).

Reporting and interpreting benchmark and sensitivity analyses

In this section, we discuss a few ways to summarize the results of the sensitivity analyses to include in a report. For each key decision, we recommend selecting one benchmark analysis and one sensitivity analysis. In all cases, the sensitivity analysis should be an alternative approach that is justifiable and appropriate. The body of the report typically presents the results from the benchmark analysis and briefly summarizes the findings from the sensitivity analyses, to show the robustness (or lack thereof) of the findings. Ideally, the appendix to a report would include tables with the results from the benchmark analyses and the sensitivity analyses side by side to facilitate comparison, for researchers interested in these details. Alternatively, if space is constrained, the report can summarize the results and note that sensitivity analyses are available upon request.

Notably, comparing and synthesizing the results of two or more analyses is not always straightforward. One challenge is that different methods can yield different estimates because of statistical noise, rather than a substantive difference between them. The following approaches enable researchers to summarize quantitatively the results of the sensitivity analyses. The most appropriate approach depends on the evaluation and type of analyses.

1. ***Create summary statistics of the differences.*** In this approach, the summary includes basic statistics like the fraction of impacts in the benchmark and sensitivity analyses that have the same direction and level of statistical significance. For example, the report might state that the impacts from the sensitivity analysis and benchmark analysis had the same sign 90 percent of the time and same level of statistical significance 80 percent of the time. This approach can be applied within a given outcome/research question or could be presented pooled across multiple outcomes (for example, different calculations for frequency of sexual intercourse and use of protection).

Consideration: This approach might be effective when there are many outcomes and discussing the difference for each outcome would be cumbersome.

2. ***Provide a range of the estimates.*** Another option is to present a range of estimates of the magnitude and statistical significance. For example, one analysis might yield an impact of 0.10 and a *p*-value of 0.03 and another might yield an impact of 0.08 and a *p*-value of 0.05. The report could give the range for the impact as 0.08 to 0.10 and the range for the *p*-value as 0.03 to 0.05.

Consideration: This approach might be particularly effective when there is more than one sensitivity analysis for a given benchmark model so that comparing the benchmark with each alternative would be challenging.

3. ***Graph the distribution of estimates for each outcome or outcome domain.*** Similar to presenting a range, researchers could graphically present the distribution of estimated effect sizes, statistical significance of the findings, or both. For example, a forest plot (Hedges and Olkin 1985), can serve to show the point estimate and confidence intervals for a benchmark and sensitivity analyses. Figure 2 shows two such examples. In the first panel focused on sexual initiation as an outcome, the benchmark result (presented in bold) shows a positive program impact (a 6 percentage point reduction in initiation rates) that is statistically significant, because the confidence interval does not cross 0. The direction of all four sensitivity analyses are all the same as the benchmark, and two of them are statistically significant. The implication from this panel is that the results look relatively robust, suggesting a positive program impact. The second panel focused on risky sex presents a somewhat different takeaway. Although the benchmark result shows a positive and statistically significant program impact, the sensitivity analyses do not consistently reproduce this result. None of the sensitivity findings is statistically significant, and in one specification (Sensitivity 4), the point estimate changes direction. This shows that the benchmark finding is not necessarily robust, and that it might actually be a spurious finding.

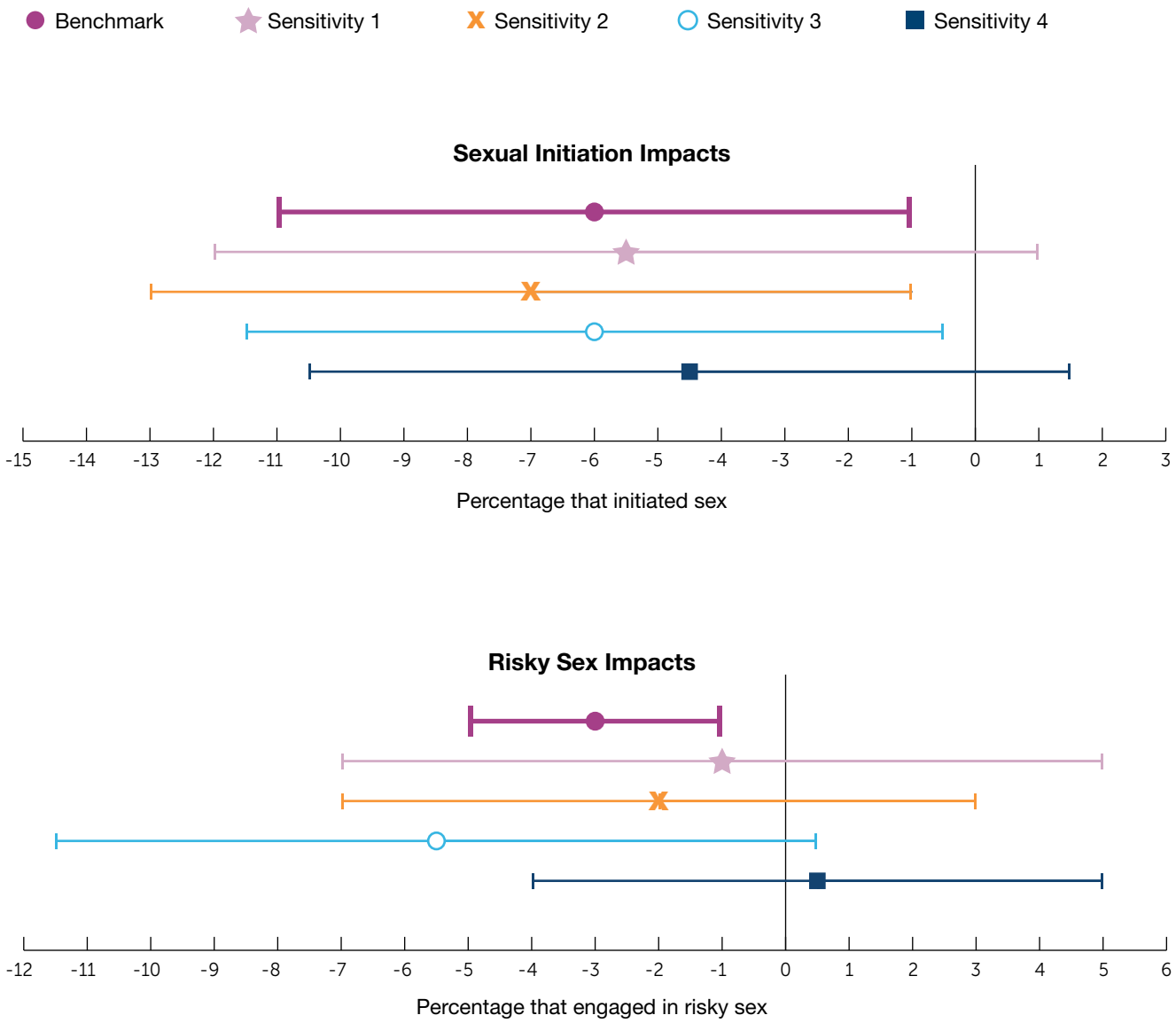
Consideration: This approach might be particularly effective when there are many sensitivity analyses for a given benchmark model so that comparing the benchmark with each alternative would be challenging and providing a range might be misleading, because there could be some outliers that drive the range.

Regardless of the approach, we recommend complementing the quantitative summary with a written interpretation of the sensitivity analysis. This interpretation focuses on the big picture implications of the research and the general consistency of findings between the benchmark and sensitivity analyses. For example, an evaluation might have many measures that capture the frequency of sexual activity. It is possible that for some measures, the statistical significance differs between the benchmark analysis and the sensitivity analysis but in both analyses the effects tend to be negative. In this case, the summary could state that the benchmark analysis and sensitivity analysis both tended to find negative impacts on frequency of sexual activity. There are, however, no hard and fast rules that can be applied to determine whether the benchmark and sensitivity analyses yield substantively different conclusions, so some judgement is required.

Although we recommend specifying the benchmark and sensitivity analyses in advance, it is possible that some evidence will suggest that one of the planned sensitivity analyses is more suitable than the planned benchmark analysis for presentation as the main findings in the text. As an example, it is possible that a variable that was planned for use as a covariate in the proposed bench-

mark model has more missing data than anticipated. Therefore, this variable might be excluded from the actual benchmark model presented in a final analysis and instead included in one of the sensitivity specifications. It is also possible that a planned method turns out not to be feasible. For example, sometimes methods based on maximum likelihood cannot be estimated because the model does not converge to a stable set of estimates. In situations like these, we recommend presenting one of the sensitivity analyses in the main text and explaining the reasons for the divergence from the impact analysis plan.

Figure 2. Graphical depictions of sensitivity analyses



Note: The symbols indicate impact estimates and the horizontal brackets indicate the 95% confidence interval of the impact estimate.

APPENDIX A.

Table A.1. Aspects of analytic approaches determined by research question

Approach	Description	Effect on analysis
Level of inference	For some types of designs, the estimates could apply either to individuals (such as students) or to groups of individuals (such as schools). For example, the study could examine whether an intervention affected sexual behavior for the average student or the average school. The focal population of the research question determines the level of inference.	This decision affects how observations are weighted. For example, if the level of inference is individuals, then all individuals would receive equal weight, but if the level of inference is schools, then the analysis would weight each school equally.
Treatment parameter	Evaluations can estimate different impact parameters that have different interpretations. For example, the Intent to Treat (ITT) parameter is the impact of being assigned to the treatment group, whereas the Treatment on the Treated (TOT) parameter is the impact of receiving treatment for those who decide to accept it. The focal population of the research question determines the treatment parameter.	Different analytical approaches yield different treatment parameters. For example, the ITT can be estimated by taking the difference in means between those assigned to the treatment and control groups. If there are no members of the control group who receive treatment, the TOT can be estimated by using random assignment as an instrumental variable for receiving treatment. A forthcoming brief will describe approaches for estimating the TOT parameter as an alternative to the ITT parameter.
Finite versus super population	Researchers must decide whether the results apply to a finite population model, in which the impact findings pertain to the study sample, or a super population model, in which the impact findings are representative of a larger sample. A super population model might be particularly relevant if the study sample was randomly (and purposefully) sampled from a larger population, such as if study schools were randomly selected from a district. The focal population of the research question informs this choice.	The choice of a finite population versus a super population model will affect how the standard errors are calculated. The exact difference will depend on the type of estimator being used, such as whether it is a design-based estimator or an HLM.

APPENDIX B

In this appendix, we focus on data preparation and analytic approaches appropriate for common situations in TPP studies, in which some observations are missing key baseline data and other observations are missing key outcome data. Table B.1 defines four types of missing data based on these categories. For example, Type A individuals have complete data on all variables of interest, whereas type B individuals have baseline data but not outcome data. Most TPP studies will include some individuals representing each of these four types because of survey or item nonresponse, and therefore, it will be important to have a strategy for addressing missing data for these different types of individuals when the goal is to present a credible impact analysis that meets HHS evidence standards.

Beyond the approaches for describing the amount of missing data presented in Deke and Puma (2013), we recommend reporting the prevalence rates of observations of types A to D. This analysis helps set the stage for the two analytic approaches necessary for credible presentations of program effectiveness: (1) a demonstration of equivalence at baseline of the analytic sample and (2) an estimate of program impacts. After demonstrating the prevalence of different types of respondents, we suggest a few approaches for these two key analyses that differ in which types of respondents are included. We do not describe these approaches in detail here for the sake of brevity.

1) ***Complete case analysis without imputation (uses only Type A respondents).*** For this approach, include Type A respondents in (1) the demonstration of baseline equivalence and (2) the estimation of program impacts.

2) ***Complete case analysis after baseline imputation (uses Type A and Type C respondents).*** Under this approach, first impute missing baseline data to a constant with a dummy indicator variable for Type C respondents. Then, demonstrate baseline equivalence of Type A individuals, and finally, estimate impacts using Type A and C individuals (after including the dummy missing data indicators as covariates).

3) ***Multiple imputation or maximum likelihood imputation (uses Types A, B, and C respondents).*** Under this approach, plan on including two separate demonstrations of baseline equivalence: (1) demonstrate the equivalence of the sample only using Type A individuals and (2) demonstrate equivalence only using both Type A and B individuals. Then, if using multiple imputation, impute missing baseline and outcome data for Type A, B, and C individuals. Finally, estimate and report impacts using imputed data (or using all observed data under maximum likelihood approach) using Type A, B, and C individuals.

Table B.1. Types of observations based on whether outcome and baseline data are missing

Type	All key baseline data observed	All key outcome data observed	Description
A	Yes	Yes	Complete data at baseline on key variables of interest that will be included in the impact estimation (for example, demographics or baseline measures of outcomes of interest) and follow-up; that is, no missing data on any key variables
B	Yes	No	Complete data at baseline but missing follow-up and outcome data
C	No	Yes	Complete data at follow-up, but missing one or more key baseline variables to be used in the analytic approach for estimating impacts
D	No	No	Missing data at both baseline and follow-up

References

Cameron, A.C., and D.L. Miller. (2015). "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317–372.

Cole, R. and R. Agodini. "Baseline Inequivalence and Matching." Washington, DC: U.S. Department of Health and Human Services, Office of Adolescent Health, 2014.

Deke, J. "Frequently Asked Questions About the Implications of Clustering in Clustered Randomized Controlled Trials (RCTs)." Washington, DC: U.S. Department of Health and Human Services, Office of Adolescent Health, 2013.

Deke, J. "Using the Linear Probability Model to Estimate Impacts on Binary Outcomes in Randomized Controlled Trials." Washington, DC: U.S. Department of Health and Human Services, Office of Adolescent Health, 2014.

Deke, J., and M. Puma. "Coping with Missing Data in Randomized Controlled Trials." Evaluation Technical Assistance Brief 3. Washington, DC: U.S. Department of Health and Human Services, Office of Adolescent Health, 2013.

Deke, J. "Design and Analysis Considerations for Cluster Randomized Controlled Trials That Have A Small Number of Clusters." *Evaluation Review*, vol. 40, no. 5, 2016, pp. 444–486.

Goesling, B. "Common Problems in Working with Data on Adolescent Sexual Risk Behaviors." Presentation to Office of Adolescent Health and Administration on Children, Youth & Family at the Teenage Pregnancy Prevention Conference, Washington, DC, 2012.

Gorsuch, R.L. *Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1983.

Hedges, Larry V., and Ingram Olkin. *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press, 1985.

Hoyle, R.H. *Handbook of Structural Equation Modeling*. New York: Guilford Press, 2012.

Imbens, G.W. "Experimental Design for Unit and Cluster Randomized Trials." Prepared for the International Initiative for Impact Evaluation, 3ie, 2011.

Imbens, G.W., D.B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge, United Kingdom: Cambridge University Press, 2015.

Kautz, T., P. Schochet, and C. Tilley. "Comparing Impact Findings from Design-Based and Model-Based Methods: An Empirical Investigation." (NCEE 2017–004). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development, 2017. Retrieved from http://ies.ed.gov/ncee/edlabs.

Leamer, Edward E. "Sensitivity analyses would help." *The American Economic Review*, vol. 75, no. 3, 1985, pp. 308-313.

Mathematica Policy Research. "Identifying Programs That Impact Teen Pregnancy, Sexually Transmitted Infections, and Associated Sexual Risk Behaviors Review Protocol Version 5.0." Retrieved from https://tppevidencereview.aspe.hhs.gov/pdfs/TPPER_Review%20Protocol_v5.pdf.

Nunnally, J.C., and I.H. Bernstein. *Psychometric Theory*, Third edition. New York: McGraw-Hill, 1994.

Raudenbush, S.W., and A.S. Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods*, Second edition. Thousand Oaks, CA: Sage, 2002.

Rosenberg, M. *Society and the Adolescent Self-Image*. Princeton, NJ: Princeton University Press, 1965.

Scher, L., and R. Cole. "Considerations Regarding Evidence Review Standards When Using RCT-YES." Houston, TX: Decision Information Resources, Inc, 2017. Retrieved from https://www.rct-yes.com/Content/PDF/Evidence%20Standards%20When%20Using%20RCT-YES.pdf.

Schochet, P.Z. "Statistical Power for Random Assignment Evaluations of Education Programs." *Journal of Educational and Behavioral Statistics*, vol. 33, no. 1, 2008, pp. 62–87.

Schochet, P.Z. "Statistical Theory for the RCT-YES Software: Design-Based Causal Inference for RCTs." (NCEE 2015–4011). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development, 2016.

White, H. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, vol. 48, no. 4, 1980, pp. 817–838.

Wasserstein, R.L., and N.A. Lazar. "The ASA's Statement on *p*-Values: Context, Process, and Purpose." *American Statistician*, vol. 70, no. 2, 2016, pp. 129–133.

This brief was written by Tim Kautz and Russell Cole from Mathematica Policy Research for the HHS Office of Adolescent Health under contract #HHSP233201300416G.