Evaluation Technical Assistance Brief
for OAH Teenage Pregnancy Prevention Grantees

September 2017

Estimating Program Effects on Program Participants

Randomized control trials (RCTs) are considered the gold standard for evaluation because they create an opportunity to calculate an unbiased estimate of the effect of offering *a program to a population of interest. In an RCT, individuals (or clusters of individuals) are randomized to either a treatment group or to a control group. The treatment group is offered an opportunity to participate in a program, the control group is not. Because the groups are formed randomly, we expect that the only systematic difference between them is whether they are offered the opportunity to participate in the program. Therefore, any systematic difference between the groups in average outcomes can be causally attributed to the offer of the program. The impact on average outcomes of the offer of the program is often called the impact of the "intent to treat" (ITT).*

The effect of offering *a program is not necessarily the same thing as the impact of* participating *in a program. Even in a well-designed RCT, study participants may not always choose to comply with their assigned conditions. Some individuals assigned to the treatment group may choose not to participate in the program and some individuals assigned to the control group might find a way to participate in the program.*

In this brief we describe (1) the research questions that are best answered by ITT impacts and impacts on participants, (2) the conceptual framework we can use as the basis for calculating a valid impact on program participants, (3) two valid approaches to estimating the impact on program participants, and (4) suggestions for how to present ITT impacts and impacts on participants in final reports or journal articles.

Research questions answered by ITT impacts and impacts on program participants

The ITT tells us the effect of the offer, which is often of interest to policy makers because the offer of a program is the policy lever they control – they typically cannot force people to participate in a program. The ITT impact should always be estimated and reported in TPP evaluations. This is the expectation of the Office of Adolescent Health (OAH) for funded grantees reporting evidence of program effectiveness.

On the other hand, the Treatment on the Treated (TOT) impact tells us the impact of program participation, which may also be of interest in contexts where it might be feasible to encourage more participation. Also, we might need to know the TOT impact in the context of a cost-benefit study, particularly if the costs are different for participants than for non-participants (that is, if costs are really driven by participants, then we need to know the benefits to participants).

Conceptual framework that supports valid estimation of the impact on participants

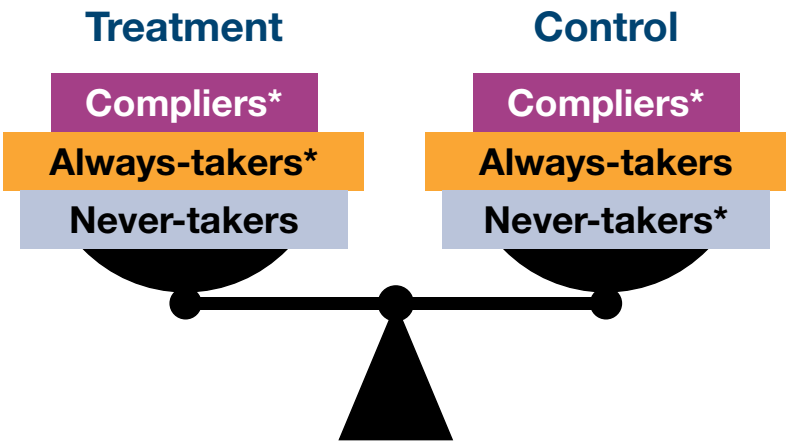
Valid estimation of a program's impacts on program participants is supported by the potential outcomes framework (Holland 1986, Rubin 1974, 2005). Under this framework, we can classify each individual in the evaluation as one of three types of individuals: (1) compliers, (2) always-takers, and (3) never-takers. Compliers are those who take up the program when they are randomly assigned to be offered the program, or do not take up the program when assigned to the control condition. Always-takers refer to people who always participate in the program, regardless of the assigned condition. Similarly, never-takers refer to those who never participate in the program, regardless of the assigned condition. Random assignment ensures that, in expectation, there are balanced numbers of these three types in the treatment and control groups (Figure 1).

Program participation definition

For the purposes of this brief, we define participation, or program take-up, as receiving any aspect of the treatment condition. It is also possible to use an alternate definition of program take-up, where participants are defined by receiving a sufficient dose of the intervention (e.g. receiving the components of the intervention defined by the program developer to constitute the minimum required dosage). This minimum required threshold may be best determined by developers familiar with the institutional details of the program. The suggested methods described in this brief can also be used to estimate the impact of treatment dosage captured as a continuous variable (see Angrist & Imbens 1995 for more details).

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

Figure 1. Composition of treatment and control groups



* represents groups of individuals who will ultimately receive the condition to which they were assigned.

Simple subgroup comparisons of those individuals who actually receive the condition to which they were randomly assigned are not rigorous approaches to estimating the TOT parameter. This is because among sample members assigned to the treatment group, the individuals who actually receive the treatment include both compliers and always-takers (this is indicated by a * in Figure 1). Similarly, among sample members assigned to the control group, the individuals who actually receive the control condition include compliers and never-takers (again, indicated with a * in Figure 1). Therefore, by comparing the subset of individuals who actually receive the condition to which they were assigned, we are comparing a compositionally dissimilar set of individuals (always takers are found in the treatment condition, but not the control condition, and never takers are found in the control condition, but not in the treatment condition). Such an analysis effectively compromises the integrity of the RCT and therefore, the evidence from such an analysis has threats to internal validity. In sum, a causally valid estimate of the TOT cannot be estimated with a subgroup analysis.

On the other hand, there are well-established ways of using the random assignment process, coupled with information on program take-up to obtain a rigorous and internally valid TOT estimate, as discussed below in Section IV. An important caveat is that an internally valid TOT estimate can be obtained and attributed to only a portion of the full randomly assigned sample – the subset of compliers. This is because the impacts of the program on those who always participate or never participate are logically zero; the program is not going to change their participation behavior in any way (that is, they either always participate or never participate). As a result of this logic, the ITT impact is driven by the subset of the sample who are compliers, and therefore, we can use this intuition to obtain an estimate of the intervention on this subset of compliers.

The effect of treatment receipt for the subgroup of compliers is referred to as the Local Average Treatment Effect (LATE), or the Complier Average Causal Effect (CACE), (note: we use LATE for remainder of this brief). The TOT is a weighted average of the treatment receipt on always-takers and compliers – and the latter component (based on compliers) is the LATE estimate that we can obtain using methods described below. Notably, when there are no crossovers from the control group (i.e. individuals randomly assigned to the control group who ultimately took up the program), then the TOT and LATE estimates are equivalent.

Analysis approaches to calculating the impact on program participants

Estimating the ITT effect is straightforward. The ITT estimate is essentially the difference between the treatment group and control group mean (often adjusted for baseline differences), regardless of the degree of compliance. This brief will not go into the details of estimating ITT impacts. See Cole et al (2013) or Kautz and Cole (2017) for further guidance on estimating the ITT parameter in the context of a TPP evaluation with the goal of meeting HHS evidence standards.

Below, we provide details on three approaches for credibly estimating the TOT parameter with the LATE, which are dependent on the types of data available to approximate compliance/take-up.[1] In the subsequent section, we focus on take-up as defined as receiving any component of the intervention.

[1] There are other approaches to estimating the TOT parameter, such as maximum likelihood or Bayesian methods (Imbens and Rubin 1997). However, the three approaches we describe are the most common ones used.

Whenever individual-level data on program take-up are available, use instrumental variable (IV) methods for estimating the LATE, i.e., effects of treatment receipt among compliers.

The IV estimator uses only the variation in take-up that is induced by the random assignment process to estimate the impacts of taking up the intervention on outcomes. It entails using random assignment indicator to predict intervention participation, and then using that predicted version of program participation as the variable that is used to show program effectiveness. The IV estimator performs well only if sample members' randomly assigned status has a strong association with take-up. The Appendix further describes guidelines for determining whether the instrument of sufficient strength to produce credible estimates.

The Exclusion Restriction Assumption

The main assumption behind the IV estimator is that neither the outcomes of always-takers nor the outcomes of never-takers differ between the interven-tion and comparison groups. As noted earlier, this is because the actual treatment or control experiences that these individuals will receive is not affected by the random assignment process – for example, an always-taker will always experience the treatment, regardless of the result of random assignment. As a result, being offered the intervention can affect outcomes only by influencing whether people enroll in the intervention. Therefore, any difference between the intervention and comparison groups must be attributable to compliers. Likewise, the difference in take-up rates between the treatment and control groups reveals the fraction of study sample members who are compliers. Conceptually, an IV estimator therefore estimates the effect of the intervention on compliers by dividing the difference in outcomes between the intervention and comparison groups by the difference in take-up rates (this is known as the Bloom [1984] adjustment, which we discuss below).

The IV estimator is usually produced using two-stage least squares (TSLS). We provide a brief description of the derivation of the IV estimator and basic steps of TSLS in the Appendix.

If individual-level data are not available, and there are no crossovers, use the Bloom adjustment to estimate the LATE:

The process for obtaining the LATE using the Bloom adjust-ment requires a series of steps:

1. Obtain an ITT estimate of program effectiveness
2. Calculate the compliance rate (C) = percent of individuals in the treatment group (among the ITT impact analytic sample) who actually took up the program.
   - The compliance rate tells us what fraction of the analytic sample belongs to the group of compliers. This means that the ITT is just the program's effect on this special group (the LATE) multiplied by the proportion of people who belong to that group (the compliance rate).
3. Simple arithmetic then gives us the LATE = (ITT estimate) / C.
   - For example, if the ITT impact is a reduction in risky sexual behavior by 10 percentage points, and only 80% of the treatment group actually attended the program, then the LATE would be 10 /.80 = 12.5 percentage points.
   - When there are no baseline covariates and no crossovers, the Bloom estimator is equivalent to the IV estimator.

If individual-level data are not available, and there is two-sided non-compliance, the following steps can be used to estimate the LATE:

If there is two-sided non-compliance, that is, members of the group assigned to the treatment who do not take up the interven-tion, AND members of the group assigned to control who do take up the intervention (i.e. crossovers), an alternate Bloom-like adjustment is possible.

- Calculate the compliance rate (C') = percent of the treatment group who were actually treated (treatment compliers) - percent of the control group who received the treatment.
- Extending the example above, if 80% of the treatment group attended the program and 20% of the control group also attended the program, then C' = 80% - 20% = 60%. Hence if the ITT estimate = 10 percentage points, then LATE = .10 / .60 = 16.7 PP.

In both cases above where LATE is derived mathematically by scaling the ITT estimate by a compliance rate (i.e. the Bloom adjustment), the standard error for the LATE estimate is typically scaled by the inverse of the compliance rate. Hence, the t-statistic (estimate divided by the standard error), as well as the p-value, for the LATE estimate using the Bloom adjustment will be identical to that of the ITT estimate. However, the standard error of the LATE estimate calculated using the Bloom adjustment will not be correct because it does not take into account uncertainty in the compliance rate (Schochet & Chiang 2009). As a result, the standard errors of the LATE estimate derived in this "division" approach are likely biased, and the associated p-values and statistical significance should be interpreted with caution.

Suggested Analytic Approach

We recommend using IV methods as a first-line approach to calculating the LATE whenever possible to ensure accurate standard errors and *p*-values.

Reporting and interpretation considerations

When planning a journal article or final report that includes a TOT result as a supplement to an ITT analysis, consider the following suggestions for completeness and transparency. First, we recommend leading with the ITT analysis as the benchmark test of program effectiveness. This is often the most policy relevant analysis, the most easily communicated result, and the impact estimate that requires the fewest assumptions for credibility. The typical reporting requirements for an ITT estimate of a TPP program have been presented in other products, and include reporting the magnitude of the effect, and the statistical significance of this impact (e.g. Murphy & Knab 2015 or the Cohort 1 impact report template).

As noted above, if a study experiences substantive non-compliance, and additional TOT analyses would be appropriate to report, the researcher will need to include the key pieces of information for critical readers and/or evidence reviewers. Evidence reviews may treat some types of LATE analyses as equally credible as ITT ones, and thus, potentially eligible for the highest evidence rating. For example, see the What Works Clearinghouse (WWC) Standards Guidance for Reviews of Studies that Present a Complier Average Causal Effect (Chapter IV of the WWC Reviewer Guidance for Use with the Procedures and Standards Handbook) for details on the requirements that this evidence review examine when reviewing the credibility of TOT estimates.

In order for LATE estimates to be viewed as credible by critical readers and reviewers, additional work will be required, above and beyond showing compliance rates and the LATE estimate. In particular, the credibility of the estimate depends on whether the exclusion restriction holds and whether sample members' randomly assigned status has a strong enough association with take-up.

For studies planning on reporting a LATE estimate, we recommend including the following information:

1. ***Report compliance rates with random assignment among the analytic samples*** – for example, the percent of the youth randomly assigned to receive the intervention, who contribute to the impact estimate, who actually took up the intervention (and comparable information for the control group).
2. ***Demonstrate baseline equivalence of the analytic sample*** – this step is probably already conducted as a part of the ITT analysis, but will be needed as part of a compelling argument for the credibility of a LATE analysis.
3. ***Report the TOT estimate and accompanying key statistics:***

a. If estimating LATE using IV methods (individual data on compliance are available):

   i. Report the F-statistic for the statistical significance of intervention offer in predicting intervention take-up.

   ii. Report standard errors of LATE estimates from the IV estimation, along with a p-value for the statistical significance of the impact estimate.

b. If estimating LATE when individual data are not available:

   i. Report how Bloom adjustment was performed, using the ITT estimate and the reported compliance rates, for transparency.

An important consideration when reporting both ITT and LATE estimates in a report or journal article is that the significance of the impacts could differ, when the recommended IV methods are used to obtain the LATE estimate. For example, it is possible that a study might have a statistically significant ITT impact but an insignificant LATE estimate, or vice versa.

Our recommendation is to focus primarily on the statistical significance of the ITT estimate, and to interpret the LATE estimate as a "rescaling" of the ITT. If the ITT estimate is statistically significant, then it is justifiable to discuss the magnitude of the rescaled LATE estimate, even if is not statistically significant. On the other hand, if the ITT estimate is not statistically significant, but the LATE estimate is, then interpretation will need to be more nuanced. In such a situation, one possible interpretation might be that the offer of the intervention did not produce a change in participant outcomes, but that the intervention did actually change outcomes among the subset of individuals who were randomly assigned to receive the program, and actually participated in the program. In doing so, we acknowledge the ITT estimate first and foremost, in terms of our interpretation of the effect of the program.

References

Angrist, J., & Imbens, G. (1991). Sources of identifying information in evaluation models.

Angrist, J. D., & Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. Journal of the American statistical Association, 90(430), 431-442.

Cole, R., Deke, J., & Zief, S. "Teen Pregnancy Preven-tion Evaluation Technical Assistance – Analysis Plan Frequently Asked Questions" Evaluation Technical Assistance FAQ. Submitted to the Office of Adolescent Health and the Administration on Children, Youth and Families Teenage Pregnancy Prevention Grantees. Princeton, NJ: Mathematica Policy Research, May 2013.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* 81(396), 945-970.

Imbens, G. W., & Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. The Annals of Statistics, 305-327.

Institute of Education Sciences. *What Works Clearinghouse (WWC) Reviewer Guidance for Use with the Procedures and Standards Handbook (version 3.0)*. Available here: https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_reviewer_guidance_030416.pdf

Kautz, T., and Cole, R. "Benchmark and Sensitivity Analyses." Evaluation Technical Assistance Brief. Submitted to the Office of Adolescent Health and the Administration on Children, Youth and Families Teenage Pregnancy Prevention Grantees. Princeton, NJ: Mathematica Policy Research, September 2017.

Murphy, L., and Knab, J. "Understanding the HHS Teen Pregnancy Prevention Evidence Review." Evaluation Technical Assistance Brief no. 8. Submitted to the Office of Adolescent Health and the Administration on Children, Youth and Families Teenage Pregnancy Prevention Grantees. Princeton, NJ: Mathematica Policy Research, June 2015

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688-701.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469), 322-331.

Schochet, P. Z. & Chiang, H. (2009). Estimation and Identification of the Complier Average Causal Effect Parameter in Education RCTs (NCEE 2009-4040). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

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

Technical appendix

A. LATE in an IV framework (Imbens and Angrist, 1991)

In a RCT, let Z_i be a dummy variable indicating random assignment to the treatment group; D_i is a dummy indicating whether treatment was actually received. In an IV framework, the IV is then Z_i, which affects treatment take-up D_i, which in turn affects the outcome of the individual, Y_i.

For simplicity's sake, we begin with an example of a simple outcomes model with no covariates, where the relationship between the outcome y and the random assignment status Z_i, is represented as

(1) $$Y_i = \beta_0 + \beta_1 Z_i + \varepsilon_i$$

where Z_i=1 if the individual was randomly assigned the program, and 0 otherwise. β_1 then measures the effect of the offer of the intervention – the ITT estimate. However, if we are interested in estimating the effect of receiving the program, the TOT estimate, then an alternate specification is required. If we simply attempted to estimate a comparable model, where we replaced treatment status Z_i with program take-up D_i, then we would have

(2) $$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$$

In this alternate specification under Equation (2), β_1 then measures the treatment effect of interest – the effect of receiving the program. However, even in an RCT, β_1 will not consistently estimate the TOT parameter, because take-up of the program is voluntary. In other words, individuals who choose to participate in the program may be systematically different than those who do not, and there will be selection bias leading ε_i and D_i to be correlated.

The random assignment of intended treatment status, Z_i, allows us to disentangle the causal effects of the program in the face of non-compliance. By virtue of random assignment, Y_i and Z_i are independent. It therefore follows that $E[\varepsilon_i | Z_i]=0$.

By taking conditional expectations of Equation (2) with Z_i alternating between 1 and 0, we can obtain the formula for the treatment effect of interest

(3) $$\beta_1 = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}$$

Thus, the causal effect of treatment receipt is given by the causal effect of treatment eligibility (the ITT estimate, β_1 from Equation 1) divided by $E[D_i \mid Z_i=1]$-$E[D_i \mid Z_i=0]$. In an RCT with no crossovers, this can be interpreted as the compliance rate in the treatment group. More generally, the denominator in Equation (3) is the difference in compliance rates by assignment status.

Key assumptions supporting LATE estimation

1. *Independence* – the instrument is independent of potential outcomes and treatments. In a well-executed RCT where the instrument is the randomly assigned treatment status, this assumption should hold true.
2. *Exclusion* – The instrument affects Y_i only through D_i. That is, the randomly assigned treatment indicator should affect the outcome of interest only through take-up of the program. The exclusion restriction is what allows us to proceed from reduced form ITT effects to TOT effects.
3. *Monotonicity* – By virtue of monotonicity, an instrument can only affect treatment take-up in one direction; that is, being assigned to the treatment group will only increase the likelihood of take-up, and never decrease it. Although this is not directly testable, this assumption should hold true in a well-executed RCT.

B. Steps to conduct TSLS

IV estimates are typically calculated using two-stage least squares (TSLS). In models without covariates, the coefficient produced by the TSLS estimator using a dummy instrument is identical to the Bloom adjustment (although the standard errors may differ). However, TSLS can also be used for models with exogenous covariates, multiple instruments and multiple treatments.

Suppose the setup is the same as in equation 2, but now we also include baseline covariates in the regression model, X_i. Under this specification, the structural equation of interest is:

(4) $$Y_i = X_i'\beta + \beta_1 D_i + \varepsilon_i$$

As its name implies, TSLS involves two steps.

1. In the first stage, treatment receipt is modeled as a function of the instrument, i.e., the treatment assignment variable, and other covariates. This regression can be written as:

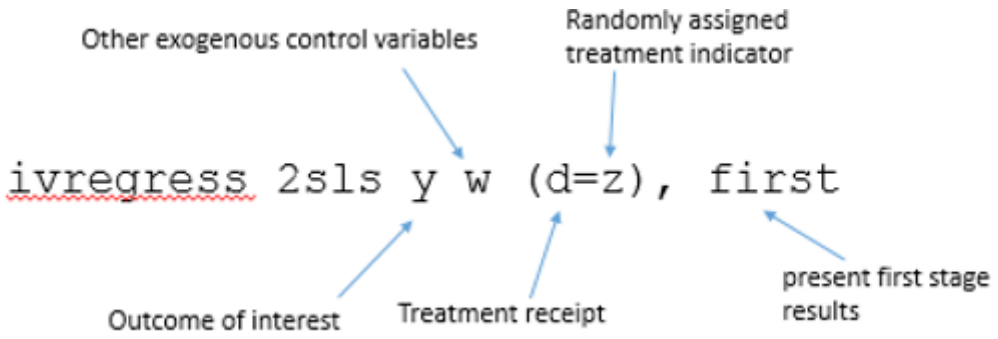
(5) $$D_i = X_i'\pi_0 + \pi_1 Z_i + u_i$$

where the coefficient π_1 is referred to as the "first-stage effect" of the instrument. The first stage equation (5) must include exactly the same exogenous covariates as appear in the structural equation (4).

- The size of the first-stage effect is a major determinant of the statistical precision of IV estimates. In an RCT, the first-stage effect approximately measures the proportion of the sample that are compliers.
- In an RCT where there is typically only one instrument (the randomly assigned treatment status), a rule-of-thumb is that the first-stage F-statistic should be greater than 10 (Staiger and Stock 1997). However, there are other more specific guidelines for determining the strength of instruments. Interested readers should refer to Stock and Yogo (2005) and refer to the WWC CACE standards for more details.

2. In the second stage, fitted values from the first-stage (predicted treatment receipt for each observation) are plugged directly into the structural equation in place of the treatment receipt indicator D_i.

- Although TSLS estimates can be constructed using these two steps, the resulting standard errors computed this way are incorrect. We therefore recommend using packaged TSLS routines in econometric software packages such as Stata or SAS.
  - Example Stata command:



The option first will report the results from the first stage regression, which will allow the researcher to gauge the strength of the instrument.

  - Ivreg2 is a similar user-written command with additional features, available via **ssc install ivreg2, replace**