How Study Design Influences Statistical Power in Community-Level Evaluations

OFFICE OF ADOLESCENT HEALTH

SEPTEMBER 2016

OAH TIER 1B EVALUATION TECHNICAL ASSISTANCE BRIEF

No. 3

OVERVIEW

There are several potential strategies that Tier 1B grantees could use to evaluate the impact of scaled-up programs, depending on the type of available data. A few that have been mentioned during Office of Adolescent Health (OAH) Tier 1B TA webinars are propensity-matched cross-sectional designs (which use data from one follow-up time point), difference-in-differences designs (which use data from one pre-intervention and one follow-up time point), and interrupted and comparative interrupted time series designs (which use data from many historical and one or more follow-up time points). The type of data available – and corresponding study design – has important implications for the sample size (i.e., number of communities) needed to detect policy-relevant impacts. This brief uses a general example to show how statistical power varies across these three types of community-level designs.

WHAT IS A POWER ANALYSIS?

A power analysis is a planning tool that evaluators should use during the study's design phase to help determine the study's likelihood of detecting meaningful impacts if they exist. If the power analysis determines that the study does not have a reasonable chance of detecting meaningful impacts, then the evaluator could take steps such as increasing the sample size (e.g., by planning to recruit additional treatment and/or comparison communities) or revising the study's design. There are two related questions that an evaluator might want to address through a power analysis:

1. Given a particular study design, a given sample size, and a given data analysis plan, what is the probability that hypothesis testing will yield a statistically significant result if the true impact of the intervention is some particular size? This probability is called the **statistical power** of the hypothesis test. (Example question: "If the intervention reduced the teen birth rate by 5 births per 1000 girls, what would be the probability that the evaluation would detect that impact?").

2. Given a particular study design, a given sample size, and a given data analysis plan, what is the smallest true impact of the intervention that can be detected with appreciable statistical power? This is called the **minimum detectable impact (MDI)**. (Example question: "What is the smallest intervention-caused reduction in the teen birth rate that can be detected with a probability of 80 percent?"). Because different evaluations of the same type of intervention may use different outcomes, MDIs can be standardized (i.e., expressed in terms of the standard deviation of the outcome) in order to make them more comparable across evaluations. A standardized MDI is called a **minimum detectable effect size (MDES)**.

Both of these questions fall under the heading of power analysis. However, many evaluators view an 80 percent probability of detecting impacts as the minimum acceptable level of statistical power. For that reason, this brief will focus on the second question; i.e., the brief will compare the MDES across various study designs holding statistical power and sample size constant.

MDE vs. MDI
The brief will focus on MDES rather than MDI because the MDI for any given evaluation depends on the base rate of the outcome, which varies considerably across grantees. For example, an MDI of 10 births per 1000 would correspond to a reduction of 50% in a community where the baseline rate was 20 births/1000, but only a reduction of 25% in a community where the baseline rate was 40 births/1000. In contrast, the MDES is comparable across communities with much different baseline outcomes.

While there is no hard-and-fast rule for what constitutes an acceptable MDES, there are some general guidelines:

- A study could be thought of as adequately powered if it has a high probability of detecting impacts of the size found in previous evaluations of similar or even competing programs (where again "high probability" is typically defined as 80 percent). If it does not, then it would be more difficult to know if the new program is similarly effective as other options..
- Alternately, a cost-benefit analysis could be used to determine the smallest impact that would be cost effective; the MDES could be set accordingly.
- A researcher could appeal to Cohen's (1986) effect sizes, in which an impact of 0.2 standard deviations is considered "small," 0.5 standard deviations is "medium," and 0.8 standard deviations is "large." Although there is some controversy about these thresholds, few evaluators would be comfortable implementing a study with an MDES of more than 0.5 standard deviations, especially in a situation where not all youth in treatment communities are exposed to the community-wide strategy. **In the data used to develop this brief's examples, an MDES of 0.5 standard deviations would translate into an intervention-caused reduction of roughly 6.7 births per 1000 teen girls, from 38.5 per 1000 prior to the intervention to 31.8 per 1000 after.** [1]

The MDES for any given study is a function of several factors, including the ratio of treatment to comparison observations, the degree to which outcomes are correlated across time or people within communities, and—crucially—the analytic strategy and sample size.

Calculating MDIs/MDESs is reasonably straightforward for an individual-level randomized controlled trial (RCT), and an OAH TA brief from the last round of TPP funding (available here) provides the necessary formulas and an Excel spreadsheet tool. Unfortunately, calculating power for a community-level quasi-experimental design (QED) is substantially more complex. For that reason, this brief explains how design choices affect statistical power—and thus required sample sizes—for various community-level QEDs, but will not provide step-by-step instructions for doing power calculations. However, the formulas used for each example in the brief are provided as an appendix. We encourage you to consult with an experienced evaluator who can perform such calculations once you have the basic outline of a design.[2]

For a community-level study, the key sample size is the number of communities in the study, not the total number of individuals in those communities. *Conceptually, because the Tier 1B strategy is community-wide, the outcomes will be measured at the community level. This is particularly true in a comparison-group design because it would be difficult or impossible to identify the comparison group at the individual level – i.e., who would hypothetically have received or benefited from the community-wide strategy if it had been offered. For this reason the sample size that feeds into your power analysis is also the number of communities rather than the number of individuals.*

EXAMPLE COMMUNITY INTERVENTION

The remainder of this brief walks through MDES/MDI calculations for various research designs using an example that illustrates key concepts. Suppose that an evaluator is asked to study a community-level TPP strategy with the following characteristics. The evaluator must determine how many communities are needed for the study:

- Treatment and comparison communities = ZIP codes
- Outcome = Teen birth rate (mean = 38.5 per 1000 female teens age 15-19)
- Data Source = Vital Statistics
- Years of Data Available = Up to 12 years total; 10 years pre-intervention and two years post-intervention

[1] In the sample of 14 California counties used to develop the brief's examples, the mean birth rate was 38.5 births per 1,000 teen girls, with a standard deviation of 13.4 births/1000.

[2] Note that the validity of a QED relies on strong assumptions that cannot entirely be appraised from the data collected. For these designs, formulas and results for MDIs/MDESs and for statistical power are invalid to the extent that the validating assumptions do not hold. In the figures we present below, we are always assuming the validating assumptions hold.

***"Intervention"** in the Tier 1B context refers to the entire strategy: community mobilization, evidence-based programming, linkages and referrals to youth-friendly health care, safe and supportive environments, trauma-informed and inclusive services.*

Throughout this example, outcome values, variances, and other statistical properties were chosen to be similar to those OAH grantees might encounter in real life. For the sake of illustration, we calculated these values using a dataset of teen births by county in California, restricting our sample to counties that had a teen birth rate higher than the national average of 26.5 births per 1000 teens age 15-19 in 2013.[3] The average teen birth rate in this sample of 14 counties was 38.5 births per 1,000. If you conduct a power calculation for your own evaluation, you should find data that you believe represent the specific communities you serve (e.g., ones with exceptionally high teen birth rates).

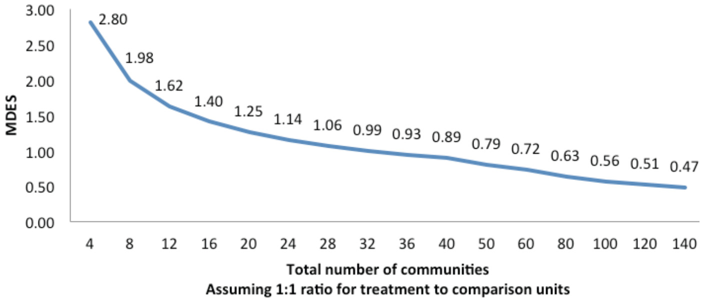
Scenario #1: Follow-up Data Only (Cross-Sectional Quasi-Experimental Design)

Suppose that historical data on the outcome of interest (e.g., teen births) are not available, meaning outcomes can be measured for the treatment and comparison groups only at a follow-up time point. In such case, the evaluator would implement what is known as a cross-sectional quasi-experimental design (QED). Typically, the comparison group would be selected using a process such as propensity-score matching on baseline characteristics. Unlike RCTs, matching designs ideally start with a large pool of potential comparison units (e.g., ZIP codes), and only the most appropriate treatment-comparison matches are retained in the analytic sample. (In practice, many or most of the potential comparison pool may go unused). Even then, it will not be possible to establish baseline equivalence on the outcome so the evaluation may lack face validity. In addition, the following example shows that community-level cross-sectional QEDs are not typically well-powered due to the absence of baseline measures of the outcome.

In a QED the comparison group will not look precisely like the treatment group, even after matching. If nothing were done to correct for differences between treatment and comparison communities, the resulting impact estimate could be biased and misleading. Fortunately, it is possible to statistically adjust for observable pre-intervention differences between the groups; e.g., using regression modelling. Whether and how this affects power depends on the details of the analysis and how good the comparison group was to begin with.

Figure 1 below shows the tradeoff between the number of communities in the evaluation (treatment plus comparison) and the MDES for a community-level QED using our sample of California ZIP codes, assuming that one comparison ZIP code is matched with each treatment ZIP code.

Figure 1. Cross-sectional QED



Note: Details of the calculation used to generate this figure are shown in the Appendix.

As Figure 1 confirms, community-level cross-sectional QEDs are not well-powered. **Even with 120 communities (60 treatment and 60 comparison), the study would not quite be powered to detect medium-sized impacts.** In our sample, a medium-sized impact of 0.5 standard deviations corresponds to an impact of approximately 6.7 births per 1000, or an intervention-caused reduction in the teen birth rate from the mean of 38.5 per 1000 to a new rate of 31.8 per 1000. Fifty communities – with 25 in the treatment group – would be required to detect even very large impacts of 0.8 standard deviations, which represents a reduction in the teen birth rate of 10.7 births per 1000 (from 38.5 per 1000 to 27.8 per 1000) in our sample. This implies that even large cross-sectional QEDs could fail to detect impacts of successful programs. It is for that reason that we strongly recommend against such a design.

Note that the MDESs in Figure 1 correspond to a design in which each treatment community is matched with exactly one comparison community. If data were available for a sufficient number of potential comparison communities, it might be possible to match more than one comparison community to each treatment community, effectively increasing the analytic sample size. Doing so will improve

[3] We converted the county-level findings to the ZIP-code level for use in the following examples using the intra-class correlation.

power – but our simulations suggest that the improvement is not dramatic for any of the designs outlined in this brief. That said, adding comparison communities will always increase power and could be worthwhile if the cost of obtaining data is low.

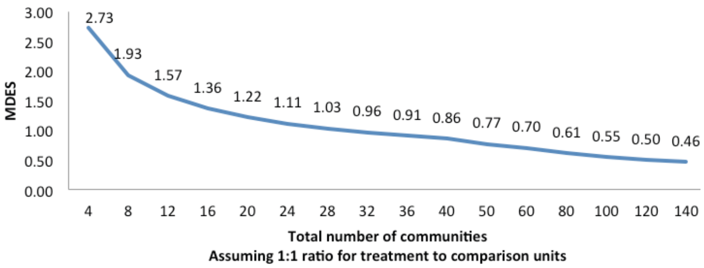
Scenario #2: Data from Baseline and Follow-Up (Difference-in-Differences / Pre-Post with Comparison Group)

Both the face validity and the statistical power of a QED can be improved by using variation over time. If the study has both treatment and comparison groups, and the evaluator is able to obtain outcome data from immediately before the intervention starts and again at follow up in each group, the evaluator might choose to implement a design called "Difference-in-Differences" (DiD).

The idea behind DiD is that (as in any QED) the treatment and comparison communities are likely to be somewhat different prior to the intervention no matter how carefully they were matched. Rather than rely on baseline demographic characteristics alone to indirectly net out these differences as in a cross-sectional QED, DiD nets out differences more directly by using the baseline measure of the outcome itself. In essence, DiD compares the change in the treatment group with the change in the comparison group, and estimates the impact as the difference between the two.

Figure 2 shows how the MDES is affected by sample size in a study using a DiD design.

Figure 2. Difference-in-differences



Note: Details of the calculation used to generate this figure are shown in the Appendix.

Comparing Figures 1 and 2 reveals that in addition to having more face validity than a cross-sectional QED, the DiD approach yields slightly better power for a given sample size (although not much in this example dataset). The reason is that the baseline outcome measure explains some of the between-community variation in the follow-up outcome measure. In this example, a DiD would require a sample of 120 communities evenly divided between the treatment and comparison groups to detect moderately-sized impacts of 0.5 standard deviations (again representing a reduction in the teen birth rate in our sample from 38.5 per 1000 to 31.8 per 1000).

Scenario #3: Data from Many Pre-Intervention Time Points (Comparative Short Interrupted Time Series)

Even with a well-matched comparison group that shows baseline equivalence on the outcome of interest (or can be made equivalent using a DiD methodology), **you may be able to further improve your power by obtaining data on the outcome of interest for several time points before the Tier 1B community intervention began.** Extant administrative data sets often include such historical data, which you can use to considerably strengthen the analysis, both in terms of face validity and statistical power, using a comparative short interrupted time series design (C-SITS).

The idea behind C-SITS is that it allows you to find a comparison group with similar baseline trends, not just similar baseline characteristics – or, more accurately, to use statistical adjustments to control for baseline differences in outcome trends.

The first Tier 1B TA webinar (which you can review here) used a graph similar to the one below to illustrate the basic concept of a C-SITS. Although C-SITS analysis can be fairly complex, this basic example shows in principle how this method can be used for analysis.

Figure 3. C-SITS

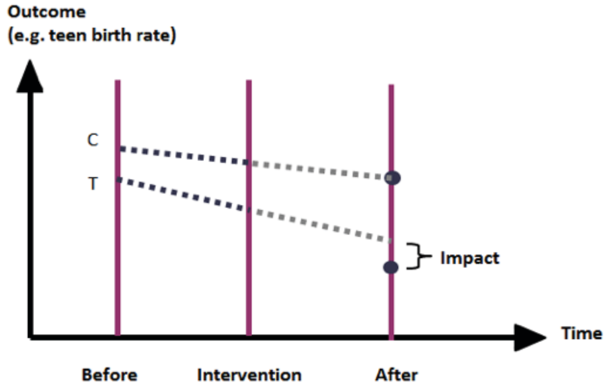
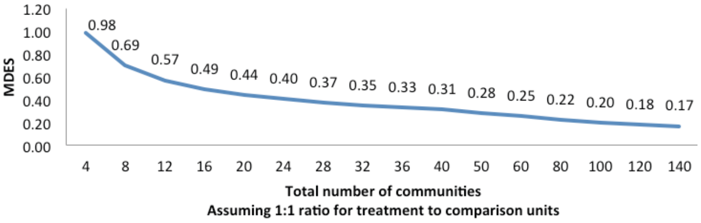


Figure 3 shows time plotted on the X axis, with the intervention occurring right in the middle of the timeline, and the outcome (teen birth rate) on the Y axis. The evaluator has obtained data for several pre-intervention time points, shown as dark blue dots, which demonstrate a trend in each group. Notice that the pre-intervention trend lines are almost, but not quite, parallel. Extending those trends into the future (light blue dots) shows what you would expect to happen in each group in the absence of the intervention. As shown, the comparison group observation happens to fall just where expected (although this will not always happen). The treatment group observation is a little better than expected. **The estimated impact is the difference between what was *expected* to happen and what actually happened.**

By incorporating even more information than the DiD, C-SITS will usually improve power compared with a DiD. Figure 4 shows the tradeoff between the number of communities in the evaluation (treatment plus comparison) and the MDES for a C-SITS that incorporates 10 pre-intervention outcome measures plus the two years of follow-up:

Figure 4. C-SITS with time trend



Note: Details of the calculation used to generate this figure are shown in the Appendix.

Figure 4 shows that to detect a small effect (using the definition of 0.2 standard deviations) with our example dataset would require 100 communities (50 treatment communities and 50 comparison communities). In the dataset used for this example, a small effect would correspond to a reduction in the teen birth rate of 2.7 births per 1000. To find a medium effect (0.5 standard deviations, or 6.7 births per 1000) would require only 16 communities (8 treatment and 8 comparison) – 104 fewer than a cross-sectional design! Because administrative data are sometimes easy to obtain for several years prior to the intervention and can substantially increase power (and face validity), we highly recommend attempting to find and use such data.

SUMMARY

Because statistical power in a community-level evaluation is driven by the number of communities in the sample rather than the number of individuals, most such evaluations will face a challenge in obtaining a large enough sample to detect moderate impacts. However, the sample size required to detect policy-relevant impacts in any particular study depends on that study's design and analytic methodology.

Therefore, evaluators can use study design as a tool to minimize this challenge, especially by obtaining administrative data on the outcomes of interest for many pre-program years.

Cross-sectional designs – those that only measure the outcome at follow up, especially cross-sectional QEDs – require very large samples to detect even moderately-sized impacts. In this brief's example, more than 120 communities would be required to detect such impacts in a cross-sectional QED. Adjusting for demographic and socioeconomic covariates in a regression model can moderately reduce these numbers. However, most evaluators are unlikely to be in this situation because pre-intervention outcome measures can almost always be obtained from administrative sources. Taking advantage of variation over time – i.e., obtaining pre-intervention measures of the outcome of interest and incorporating them in the analysis – can substantially reduce the required sample size compared with a cross-sectional design. As can be seen by comparing a difference-in-differences strategy (Figure 2) with a comparative short interrupted time series strategy (Figure 4), the more pre-intervention years for which outcome data can be obtained the better. In the brief's example, using 10 pre-intervention years' worth of data in a C-SITS reduces the required sample to only 16 communities – which is nearly an order of magnitude less than a cross-sectional QED.

APPENDIX

This appendix provides the formulas and assumptions used to produce Figures 1, 2, and 4. It is meant for a technical audience and therefore provides little descriptive explanation. Parameter values were calculated using a dataset of 14 California counties with birthrates higher than the national average in 2013, and (where appropriate) converted to ZIP-code level parameters using an Intra-Class Correlation (ICC) obtained from similar data. For each figure, the minimum detectable impact (MDI) was calculated as

1. $MDI = (t_{\alpha/2} + t_{\beta})SE(\hat{\beta}_1)$

where $t_{\alpha/2}$ and t_{β} are quantiles from a *t*-distribution, $\hat{\beta}_1$ is the coefficient on the treatment indicator in a regression model (i.e., the impact estimate), and SE $(\hat{\beta}_1)$ is the standard error of the impact estimate or equivalently $\sqrt{Var(\hat{\beta}_1)}$. To obtain a minimum detectable effect size, we divide the MDI by a standardization factor equal to the standard deviation of the outcome at the community (i.e., ZIP-code) level; in this case 181.44.

Figure 1: ANCOVA (One year of data available at follow up)

The calculations underlying Figure 1 presume that the evaluator will estimate impacts using a regression model of the following general form:

2. $Y_k = \beta_1(TrtGrp_k) + \sum_{p=1}^{P} \varphi_p X_p + \varepsilon_k$

where:

Y_k is the outcome of interest for community *k* (e.g., teen birth rate)

$TrtGrp_k$ = 1 if community *k* is a treatment community, 0 if comparison community.

X_p are other model covariates (e.g., community-level demographics).

ε_k is a residual for community *k*, assumed distributed $N(0,\sigma^2)$

The variance of the treatment effect, $Var(\hat{\beta}_1)$, yielded by model 2 is:

3. $Var(\hat{\beta}_1) = \frac{\sigma_Y^2(1-R^2)}{NP(1-P)}$

where:

σ_Y^2 is the variance of the outcome measure in the follow-up year for all treatment and comparison communities = 181.

R^2 is the net proportion of outcome variance explained by the predictor variables. We conservatively assume this is equal to zero in the absence of pre-intervention data.

N is the pooled treatment + comparison sample size (number of communities), which is allowed to vary in Figure 1.

P is the proportion of communities that are in the treatment group; we set this equal to 0.5.

Figure 2: Difference in Differences (One year of pre-intervention data; one year of follow-up data)

The calculations underlying Figure 2 presume that the evaluator will estimate impacts using a regression model of the following general form, using one year of pre-intervention data and one year of follow-up data:

4. $$Y_{jk} = \beta_1\left(TrtGrp_k * TrtYr_{jk}\right) + \beta_2\left(TrtYr_{jk}\right) + \sum_{k=1}^{K} \alpha_k ZIP_k + \sum_{p=1}^{P} \varphi_p X_p + \varepsilon_{jk}$$

where:

Y_{jk} is the j^{th} observation on community k,

$TrtGrp_k$ = 1 if community k is a treatment community, 0 if comparison community

$TrtYr_{jk}$ = 1 if observation in year *j* is post-treatment, =0 if pre-treatment

X_p are other model covariates

ZIP_k are fixed dummy variables for communities

ε_{jk} residual for j^{th} observation on community *k*, assumed distributed $N(0, \sigma^2)$

The variance of the treatment effect, $Var\left(\hat{\beta}_1\right)$, yielded by model 4 is:

5. $$Var\left(\hat{\beta}_1\right) = \frac{\sigma_Y^2(1 - \left(R^2_{Y|TG*TYr} + R^2_{Y|ZIP(TG*TYr)} + R^2_{Y|TrtYr(ZIP,TG*TYr)} + R^2_{Y|X(TrtYr,ZIP,TG*TYr)}\right))(AC)}{N\bar{T}(1 - \bar{T})(1 - R^2_{TG*TYr|ZIP} + R^2_{TG*TYr|TrtYr(ZIP)} + R^2_{TG*TYr|X(TrtYr,ZIP)})}$$

where:

σ_Y^2 is the variance of the outcome measure (all years, all treatment and comparison communities) = 468.

$R^2_{Y|TG*TYr}$ is the proportion of variance of the outcome explained by the predictor variable. Because the evaluator is testing the null hypothesis that the treatment effect is zero, this equals zero at the design phase.

$R^2_{Y|ZIP(TG*TYr)}$ This is the proportion of total variance that is accounted for by adding dummy variables for communities to the model. This is the semipartial r-squared for communities.[4] In the design phase, when investigators are conceptualizing $R^2_{Y|TG*TYr}$ as being equal to zero they can conceptualize $R^2_{Y|ZIP(TG*TYr)}$ as being the proportion of total variance that is between-communities, while the remaining variance can be conceived of as variation over time within-communities. In our data, this value is = 0.75.

$R^2_{Y|TrtYr(ZIP,TG*TYr)}$ This is the proportion of the variance in the outcome that is explained by adding the *TrtYr* variable to the model that already includes the predictor variable and community fixed effects. At the design phase, this term is also assumed equal to zero.

[4] The terminology "semipartial r-squared" comes from Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, Hillsdale NJ.

$R^2_{Y|X(TrtYr,ZIP,TG*TYr)}$ This is the proportion of the variance of the outcome measures that is explained by adding any remaining terms (e.g., covariate Xs) to the model that already includes the predictor variable, community fixed effects and the indicator for post-treatment years. This term is equal to the r-squared from the full model in Equation 4 minus the r-squared from a smaller model that does not include the other covariates. For simplicity, we assume this value is = 0.

AC is a design effect for autocorrelation. If there is autocorrelation present in the data it will inflate the variance of the treatment effect. For simplicity, we assume that there is zero autocorrelation and therefore that the design effect for autocorrelation is equal to 1.

N is the pooled treatment + comparison sample size (number of communities) across time, which is allowed to vary in Figure 2.

$\bar{T}$ is the proportion of the observations for which the treatment indicator equals one.

Specifically, for the impact model shown in Equation 4, it is the proportion of observation where $TrtGrp_k * TrtYr_{jk}$=1 ;because assume a balanced design, this = 0.25.

$R^2_{TG*TYr|ZIP}$ This is a measure of the squared correlation between the predictor variable and the community fixed effects (or indicators). We have calculated this value as 0.33.

$R^2_{TG*TYr|TrtYr(ZIP)}$ This is a measure of the squared correlation between the predictor variable and the term that indicates post-treatment observations, *TrtYr* conditional on the community indicators. We have calculated this value as 0.33.

$R^2_{TG*TYr|X(ZIP,TrtYr)}$ This is a measure of the squared correlation between the predictor indicator and the any remaining terms (e.g. covariate Xs) conditional on community fixed effects and the indicator variable for the post-treatment years. We use a value of zero.

Figure 4. C-SITS (Ten years of pre-intervention data, two years of follow-up data)

When many years of pre-intervention data are available, the evaluator could estimate impacts using the following regression model, which is similar to equation 5 but adds a pre-intervention time trend:

6. $Y_{jk} = \beta_1\left(TrtGrp_k * TrtYr_{jk}\right) + \beta_2\left(Time_j\right) + \beta_3\left(TrtYr_{jk}\right) + \sum_{k=1}^{K} \alpha_k ZIP_k + \sum_{p=1}^{P} \varphi_p X_p + \varepsilon_{jk}$

where terms are defined as in equation 4, and in addition:

$Time_j$ is a variable indicating the year *j* to allow for a time trend.

The variance of the treatment effect from this model specification is given by:

7. $Var\left(\hat{\beta}_1\right) =$

$$\frac{\sigma_Y^2(1-\left(R^2_{Y|TG*TYr} + R^2_{Y|ZIP(TG*TYr)} + R^2_{Y|TrtYr(ZIP,TG*TYr)} + R^2_{Y|Time(TrtYr,ZIP,TG*TYr)} + R^2_{Y|X(Time,TrtYr,ZIP,TG*TYr)}\right))(AC)}{N\bar{T}(1-\bar{T})(1-R^2_{TG*TYr|ZIP} + R^2_{TG*TYr|TrtYr(ZIP)} + R^2_{TG*TYr|Time(TrtYr,ZIP)} + R^2_{TG*TYr|X(Time,TrtYr,ZIP)})}$$

where all terms in common with equation 5 are defined equivalently, with the following exceptions and additions:

σ_Y^2 the variance of the outcome measure (all years, all treatment and comparison communities) has a value of 975.

$R^2_{Y|Time(TrtYr,ZIP,TG*TYr)}$ This is the proportion of the variance in the outcome that is explained by adding the *Time* variable to the model that already includes the *TrtYr* variable, the predictor variable and community fixed effects. We have calculated this value as 0.15

$R^2_{Y|X(Time,TrtYr,ZIP,TG*TYr)}$ This is the proportion of the variance of the outcome measures that is explained by adding any remaining terms (e.g., covariate Xs) to the model that already includes the predictor variable, community fixed effects and the indicator for post-treatment years.
This term is equal to the r-squared from the full model in Equation 4 minus the r-squared from a smaller model that does not include the other covariates. For simplicity, we assume this value is = 0.

$R^2_{TG*TYr|ZIP}$ This term has a value of 0.0909.

$R^2_{TG*TYr|TrtYr(ZIP)}$ This term has a value of 0.4545.

$R^2_{TG*TYr|Time(TrtYr,ZIP)}$ This is a measure of the squared correlation between the predictor indicator and the *Time* variable, conditional on community fixed effects and the indicator variable for the post-treatment years. We use a value of 0.091.

$R^2_{TG*TYr|X(Time,TrtYr,ZIP)}$ This is a measure of the squared correlation between the predictor indicator and the any remaining terms (e.g. covariate Xs) conditional on community fixed effects and the indicator variable for the post-treatment years. We use a value of zero.

Authors

This brief was written by Randall Juras, Alison Comfort, and Ed Bein from Abt Associates for the HHS Office of Adolescent Health under Contract No. HHSP233201500069I, Task Order No. HHSP23337003T