Technical Assistance Webinar
Designing Community-Level Evaluations
Tier 1B Grantees and Evaluators

Moderator: Alexandra Warner
February 4, 2016
12:00 pm CT

Coordinator: Thank you for standing by. At this time, all participants are in a listen only mode. After the presentation, we will conduct a question and answer session. To ask a question, please press the star 1 and please record your name.

Today's conference is being recorded. If you have any objections, you may disconnect at this time.

I would like to introduce your host for today's conference, Amy Farb. You may begin.

Amy Farb: Thank you very much. Greetings everybody. Welcome to our first TA webinar for OAH's TPP Tier 1 Grantees and Evaluators. This webinar is tilted Designing Community Level Evaluations. I'm Amy Farb. I'm the evaluation specialist at OAH. We thank you for joining us today. We know you're all in the throes of designing your Tier 1B projects and the evaluation is a good part of that. So we hope this webinar is timely and useful for you.

Okay, this slide shows you who's speaking to you today. We should all look pretty familiar to you. We're the same people who spoke to you about the federal evaluation at the Tier 1 orientation back in November. We have Drs. Kim Francis and Randall Juras from Abt Associates, and that's me on the end. At this point, I'm going to turn it over to Kim. She'll give you the overview for the webinar today, and then I'll speak to you again in just a little bit. Kim?

Kim Francis: Thanks, Amy. So our talk today will begin with a brief overview of Abt Associates role with the Tier 1B program and then Amy will talk a little bit about the evaluation requirements and expectations. And then the bulk of this afternoon will be with Dr. Randall Juras, who is going to guide us through some best practices for selecting comparison groups when working at the community level, and also provide an introduction to how to do this in practice using matching techniques. And then we'll finish by illustrating a good comparison group design and also going over some potentially problematic designs as well.

You should all have a Q&A box on your screen where you can type in questions to us during the webinar, and we will plan to pause a couple of times during the presentation just to check if there are any pressing questions that have been submitted through that box. We may not be able to answer all of them, but please know that we're registering all of your questions and we're planning to make answers available to all of your questions after -- at some point after the webinar. And also, after the presentation, we're planning to have a few minutes at the end where the phone line, you can open up your phone line and ask a question directly.

So we're hoping that by the end of the session, you'll understand OAH's expectations for evaluation and you'll know what constitutes a good comparison group for your community level projects, and that you'll also

understand a range of design options that can be used for evaluating community initiatives and be able to identify potentially problematic designs as well.

So briefly about our role at Abt Associates. We've provided both federal and local evaluation services for the TPP program since its inception in 2010 and we're very excited to be the evaluation design contractor for the Tier 1B program. In this role, we'll be learning from you about the scale up strategies you're using and describing them across the 50 grantee projects. We'll also be designing an evaluation aimed at evaluating the effectiveness of the overall Tier 1B grant strategy, which will involve up to 10 grantees. And we'll provide evaluation TA, including group TA like today, and on-on-one TA of the grantees that are participating in the evaluation. We expect to reach out to potential evaluation participants starting this May. And now, I'll turn it back over to Amy.

Amy Farb: All right. So if everyone will refer back to their funding opportunity announcement for the Tier1B grants, you'll recall that the evaluation requirements for this grant category include three components, performance measures an implementation study, and an impact study. The performance measures are the ones that were provided to you by OAH. The web-based system that collects the performance measures is up and running, and you should all be registered in it by now. You'll report your measures to us twice a year and if you need to administer a survey to collect them, we do have an OMB clearance number and expiration date that you can use on your surveys. And I've listed that on the slide for your use.

The second evaluation requirement is the implementation evaluation. OAH has asked you to document the development and implementation of your Tier 1B project, identifying the key successes, challenges, and lessons learned

from those activities. We expect that to do that, you'll need to hold focus groups and participant interviews with the key stakeholder groups in your project. That's your partners, the community advisory group, the youth leadership council, program facilitators, participants, et cetera.

I use the word project because this effort should encompass everything you're doing as part of your Tier 1B grant activities. It's not solely a study of the implementation of the evidence-based programs. By the end of the grant period, you'll need to submit this evaluation in the form of an implementation study report to OAH.

The third evaluation requirement and the focus of today's webinar is the impact evaluation. For the impact evaluation, OAH has asked you to identify goals, in other words, research questions for your projects. You should specify what your project is intended to affect. Is it a reduction in teen births, a reduction in STIs? Is it increased referrals for reproduction health clinics? It might have to do with academic outcomes or other outcomes. We're very flexible in that respect, but we would like you to try to focus at least one of your research questions on one of the evidence review outcomes.

We want you to determine how effective your project was in meeting that goal. We want you to address the research question using existing data to quantify the effect and demonstrate that it was due to your Tier 1B project. And by the end of the grant period, you'll prepare your findings in an evaluation report and submit it to OAH. There are a few important things to remember, though, in this. You must complete all three of these evaluation activities within 10% of your grant funds each year. This is the budget cap on these activities. So that's why we encourage you to use existing datasets. You don't have enough funds to properly track, sample, and conduct follow-up surveys with them in a rigorous manner.

Your plan or design for these evaluation activities must be approved by OAH by June 30 as part of your year one milestones. I'll repeat that. June 30 you have to have approval from OAH to have met your year one milestones. It's important to understand that I said plan. You don't have to have identified your comparison community yet. You don't have to have assigned MOU for obtaining data or have obtained the data, started cleaning it, tried merging it, et cetera. We want your plan for these activities and for that plan to demonstrate you'll be able to meet the OAH requirements for the impact evaluation.

So have a look at what the existing data is. Know the criteria for obtaining that data. Know what time period the data pertains to and speak to that. So you can also properly time your analysis activities. Propose some potential comparison communities and familiarize yourself with the potential data in those communities to know if this is really going to work. We want a well thought out plan approved by June 30.

Now, to get an approved plan by June 30, you'll need to start right now. Some of you submitted your evaluation plans with your progress report. Thank you very much for that. For others, I've already had some of your project officers reach out to you about revising or updating your evaluation plans. You'll need to start submitting your plans to your OAH project officer as soon as possible to receive that approval by June 30. Please do not submit any evaluation plans after June 1 for approval. We just simply won't have enough time to get them approved by the 30th. So please speak to your project officers if there's any confusion, any questions about this, and we're happy to get back to you guys.

The webinar today should help you in thinking through the impact evaluation design and the necessary components of it. So with that in mind, we'll turn it to Randall, who will begin with today's content.

Randall Juras: Thank you, Amy. Hi, everyone. This is Randall Juras with Abt Associates. Today, I'll be talking with you about the basic principles and key considerations for how to select a comparison group for a community level impact evaluation. As Kim mentioned, I will be pausing for questions at a couple of points, just to answer one to two questions. So if you have any questions as I go along, please type them in and Kim will select a couple for me to answer. I will also be taking both written and verbal questions after the end of the presentation.

So the purpose of a program impact evaluation is to estimate how effective the specified program was at changing a given outcome relative to what the outcome would have been if the program had not been implemented. After delivering program services, it's possible to know -- by measuring it -- what the outcome was with the intervention. That's because you did implement the intervention and then measured what happens in your sample. On the other hand, you do not know and cannot measure what would have happened in your sample if you had not implemented the intervention. We call this hypothetical state of the world. In other words, what would have happened in the absence of the intervention is the counterfactual.

The goal for researchers is to convince a skeptical reader that the intervention caused the observed impact. In the comparison group design, this is accomplished by measuring what actually happens in the intervention communities, which we call the program or treatment group, and comparing it with what happens in a group of similar communities that closely resembles the counterfactual, which is what we call the comparison group. If you want to

persuade readers that you have established causality, it's critical that the comparison group look as close as possible to how the treatment group would have looked if it had not received the intervention. In other words, the comparison group should look like the counterfactual.

Broadly speaking, this can be accomplished in two non-exclusive ways, in design and/or analysis. By design, I mean finding a comparison group that looks as similar to the treatment group as possible so that little or no additional work is needed to convince the reader that the comparison is meaningful. By analysis, I mean finding a comparison group that looks something like the treatment group and using statistical methods to adjust for as many observable differences as possible between the two groups.

I do want to emphasize the last bullet on this slide. Because statistical methods can only control for observable and measured differences between the two groups and not unobservable or unmeasured ones, I'd like to say that design trumps analysis. Hopefully, by the end of the presentation you'll have a pretty good understanding of why I think that's true. But the basic idea is that you should put as much effort into finding a good comparison group as you can and not overly rely on statistical methods.

Now, I'm going to turn to some of the key considerations of selecting a comparison group. Selecting a good comparison group begins, believe it or not, with the specification of your study's research questions. Without knowing exactly what the research question is, it is not possible to select an appropriate comparison group. And I hope it will become clear why this is true. A well-specified research question includes the four elements that you see listed on this slide. First, you need to specify what is the intervention that you're testing. For your studies, OAH expects that you'll be evaluating the entirety of the intervention, or at least as much of it as can be tested, rather

than one single component, or a couple of components, like one of the evidence-based programs, or EVIs.

In addition to the EVIs being implemented, for example, your intervention might include community mobilization efforts, health referrals, or community advisory groups. Your evaluation should be testing the impact of this whole OAH funded effort, if possible. Second, the research questions should include a precise description of the target population. In other words, for whom do you expect the program to be effective? Without knowing this, you do not know for whom you should measure outcomes. In general, my understanding is that OAH expects your programs to affect communities broadly rather than affecting only program participants who participate in one of the EBIs. So the more specific you are about defining this population, the better prepared you will be to find an appropriate comparison group and the better the reader will be able to understand to whom your results should generalize.

The third essential component of a well-specified research question is a description of the counterfactual condition or alternately, if you selected one, what the comparison group is. Typically, you won't be able to measure the impact of the program relative to nothing or no services because there essentially isn't such a thing as an untreated group in this context. Instead, you should try to understand what services would be available in the absence of your intervention, which is something we call business as usual, and make clear to the reader that you're estimating the impact of the program compared to whatever that is. It's not enough to vaguely say you're measuring an impact. You need to say what you're measuring it compared with.

Finally, the research questions should specify the outcome or outcome domain that you're interested in, which ideally would come from the logic model. And outcome domain, by the way, is simply a description of a broad construct into

which you might group several outcomes. So for example, you might group outcomes like condom use and multiple partners into the domain of sexual risk behavior. Tier 1B programs could be intended to effect pregnancy rates, birth rates, STI rates, sexual risk behavior, or even academic outcomes like graduation. Of course, you will need to find a comparison group for which you can measure whatever outcomes you specify. If you can't find data on the outcome, you can't ask that particular research question.

So here is a quick example of what I believe to be a well-specified research question, including these four components. This is for an evaluation of a hypothetical community wide program called the Abt Community Project. The question is does the Abt Community Project affect the birth rate for girls ages 14 to 19 in Cambridge, Massachusetts compared with girls in similar communities that do not have community wide teen pregnancy prevention programs? The question here defines the intervention, which is the Abt Community Project as a whole, including all of its components. It also precisely defines the target population, which is girls ages 14 to 19 in Cambridge, Massachusetts.

If you wanted to be even more precise, you might specify whether you're interested in girls, who are ages 14 to 19 at the time of your follow-up or at the time the intervention was actually implemented. Third, the question includes a description of the counterfactual condition. Actually, in this case, it includes a description of the desired comparison group, which is girls in similar communities that do not have community wide TPP programs. Notice that this does not rule out that such communities could have some TPP programming for their kids, just that they shouldn't have a comprehensive community wide effort.

Another acceptable way of specifying the counterfactual in this research question would be to describe it as being compared with what would have happened for this group in the absence of the intervention. And let me repeat that, compared with what would have happened for this group in the absence of the intervention. And notice that for this to be credible, you would need to know what teen pregnancy prevention programming would have existed in the absence of your intervention in Cambridge, Massachusetts so that you could identify a convincing comparison group that included that kind of programming.

Finally, this research question includes a description of the outcome of interest, which in this example is the birthrate. If you have more than one outcome or domain of interest, you should write an additional research question for each outcome. This will also make it apparent if you have so many research questions that you risk finding spurious results simply by virtue of asking so many questions, which is something we often see. If you have 100 outcomes, it's very likely that you'll find something to be statistically significant among those outcomes. And in that case, you might consider narrowing down the number of research questions or specifying some of them as being of primary importance.

I would like to take a second to illustrate why it's important to clearly and precisely specify the target population, both to ensure a correctly aligned comparison group and to choose a comparison that OAH finds appropriate. This slide shows two different research questions. The first research question, which specifies the target population as girls ages 14 to 19 in Cambridge, Massachusetts is asking about the community level impact of the program. The second question, which specifies the target population as girls ages 14 to 19 who participated in the intervention is asking about the impact of the

program on only those members of the community who are directly affected by participating in one of the evidence-based programs.

In both cases, it is critical to specify a comparison group that approximate the counterfactual. In other words, one that looks like the treatment group would have in the absence of the intervention. In the first question, that's all girls' ages 14 to 19 in similar communities and in the second its girls in similar communities who would have participated if given the chance. The second question is of less interest to OAH because it asks about a component of the intervention, the EBI, rather than the whole intervention. It also requires the researcher to put a lot more work into finding an appropriate comparison group because the researcher would have to define and measure eligibility criteria and model willingness to participate in that group, which can be difficult.

With the research question defined, I'd like to offer some advice on how to select a comparison group that most resembles the counterfactual. This slide shows a few of the key characteristics that you should be looking for. The first two key characteristics of a good comparison group, which we lumped together because they sound nice together, are that it should be local and focal. A local comparison group is geographically close to the same locale as the treatment group. In other words, if you're implementing a TPP program at communities in the Northeast, you would want to compare it to other communities in the Northeast, not to communities in the Southwest, which is a part of the country that might be experiencing much different economic and social trends.

A focal comparison group is one that looks similar to the treatment group in terms of other observable characteristics, when those characteristics are measured at baseline or before the intervention begins. These could be things

like demographics and the levels or trends in the outcome. The communities that you selected probably have high rates of teen pregnancy and other adverse outcomes. That's why you selected them. So you would want to select a comparison group that has long had those same adverse outcomes, like low teen pregnancy rates to represent the counterfactual. If you don't do that, the playing field probably wouldn't be level for comparing those two groups.

Another important characteristic of the comparison group is that outcomes should be measured at the same level of aggregation as in the treatment group. So if you're measuring outcomes for individual kids who participate in an evidence based program in the treatment group, which we don't recommend, it would be inappropriate to compare those outcomes with community level aggregates and communities that did not implement the program, and vice versa because those -- would include those communities would include kids who did not participate in an evidence based program.

On a similar note, outcomes should be measured in the same way in the treatment and comparison groups, which means that you need to find a common source of data across the treatment and comparison groups. There are well-known differences between surveys and administrative data, and even between different kinds of administrative data and the kinds of things that are (castered) and for whom. In many cases, you could go to a sample at a single point in time, measure the outcome using two different data sources for that one sample, and find an impact even though there isn't any real difference to measure because it's the same people at the same time. You don't want this kind of thing to bias your results. So it's good to just find a single data source that measures the outcomes across your treatment group and your comparison group.

So now, I'd like to walk through a straightforward example of how to apply these lessons to a real research question, albeit one which is not at all related to teen pregnancy on purpose, because I want to abstract from the specific problems of teen pregnancy. Suppose for a minute that you're the manufacturer of a new fertilizer and you want to answer the research question that you see here. Does applying this new kind of fertilizer to apple trees improve the seed count in apples compared with using no fertilizer? And I should emphasize that this is a hypothetical example because I know almost literally nothing about agriculture.

Notice that the research question includes all four of the key components that I talked about earlier. It specifies the intervention, which is new fertilizer. It specifies the target population, which is apple trees. It specifies the counterfactual, which is no fertilizer use, rather than using an existing fertilizer, and it specifies the outcome domain, which is the number of seeds per apple.

So how could you go about answering this question using a comparison group design? Well, the first very appealing option would be to use a randomized controlled trial, or RCT. In this design, you could select, for example, several Macintosh apple trees planted in the same soil near to each other in the same orchard. And I'm going to assume for a minute that fertilizer isn't spread in the wind so the treatment can be localized to each tree. So you paint it on, for example.

Then you could randomly select some of the trees in the orchard to treat with a new fertilizer and leave the other nearby trees untreated. At the end of the growing season, you would collect the apples, measure the number of seeds in each apple, calculate the average number of seeds per apple in each group, the treated group, and the untreated group, and then compare the averages across

the two groups to see which did better. This is a convincing comparison. Because you randomly selected the trees that comprise each group, the trees in the comparison group should look just alike, except for the fact that one group got the treatment. The comparison group here is both local and focal, but even in this case, look at that bottom bullet, the comparison will only be convincing if you have a number of trees in each group, a number that's bigger than one, just in case, for example, one tree were hit by lightning or cost a disease, which could hurt your comparison.

Unfortunately, there are lots and lots of situations in which it is impossible, infeasible, or simply too costly to implement an RCT. In fact, OAH expects that few or none of the Tier 1B grantees will be able to implement a community level RCT. In this example, you might think that, for example, maybe orchard owner already has a contract to apply the new fertilizer to all of the trees in one orchard. So what can you do?

Well, one good option would be to use a quasi-experimental design, or QED, with a comparison group, which is what we expect most Tier 1B grantees to do. So on this slide is one example among many possibilities of how a QED could be implemented, and this is an opportunity to be creative. So say that you have two neighboring orchards. Orchard One has mostly Macintosh trees with a few Golden Delicious trees in it, while Orchard Two has mostly Golden Delicious Trees with a few Macintoshes. One thing you could do is to treat all of the apple trees in Orchard One with a new fertilizer, leave all of the nearby trees in Orchard Two untreated. Just like before, you could calculate the average number of seeds per apple in each orchard at the end of the growing season and compare the averages to see which group did better.

But note that this design is less convincing than the RCT. The comparison group is reasonably local because it's in the neighboring orchard, which would

presumably be subject to the same light rainfall and other growing conditions. But it is not very focal, as the comparison group apples are largely from a different type of tree. They're mostly Golden Delicious trees in the comparison group instead of mostly Macintosh trees. Fortunately, this isn't the end of the story. You can use analytic techniques to make the comparison more convincing. Essentially, by comparing the Macintoshes in Orchard One with the Macintoshes in Orchard Two, comparing the Golden Delicious apples in Orchard One with the Golden Delicious apples in Orchard Two, and then averaging across these two differences instead of comparing the orchard wide averages with each other. This is just statistical adjustment and in this case could be implemented using a simple regression model.

Now, say that you don't have any good local comparison group but you do have a nationwide measure of the outcome of interest. One thing that you could do in that case is to benchmark your orchard that you're treating through the national average. So you treat a sample of Macintosh apple trees with the new fertilizer. You measure the average number of seeds per apple in this sample. Then you go out to some data source to find the national average seed count per apple for apple trees using national data and compare across groups. This is even less convincing than the QED was because your comparison group is likely not focal, because the national average could include lots of trees that aren't the same kinds of trees as are in your orchard. And it's also not local. The national average includes trees that are in much different growing conditions than your trees. That makes it more difficult to statistically adjust the results, although there are things that you could do.

Perhaps, for example, you could measure the difference in seed count between your treated trees and the national average before you apply the treatment, and then do it again after the end of the growing season, after you've applied the treatment so that you could see if your trees improve relative to the other

trees. This is a design called difference indifferences. But this design is still less convincing because it assumes additional things. It assumes in this case the changes over time would be the same in both types of apple, in whatever apples are in the national average, and then your local trees if the national average trees -- if your local trees were left untreated. So it's an okay option but it's somewhat less convincing.

Now, here's another design you could use that as you can probably tell from the title, I wouldn't recommend. This is a QED with an inappropriate comparison, just to illustrate the importance of the research question. In this kind of design, you would treat Macintosh trees with the new fertilizer, treat nearby Macintosh trees, for example in a neighboring orchard with the existing fertilizer, and then calculate the average number of seeds per apple in each group of trees and compare to see which did better. This comparison group is almost perfect. It is both local and focal. It's aggregated to the same level, which is the tree. You're using the same data to measure each group. You're going on accounting but note that it answers a different research question. You cannot tell by using this comparison group what the impact is compared with no fertilizer, which was your research question. You can tell what the impact was relative to using an existing fertilizer.

So if this is our only option for a comparison group, you might have to revise your research question to make it clear to your reader what the impact that you're measuring is of. And finally, I want to describe another type of design that I'll call the apples to oranges comparison. This is where you would treat Macintosh apple trees with your new fertilizer, leave far away orange trees untreated. They would have to be far away because apple trees and orange trees don't grow in the same places. You could calculate the average number of seeds per fruit in each group of trees and compare it.

Now, there's a reason that we have a common saying about comparing apples to oranges. This comparison group is neither local nor focal, and this case there is no amount of statistical adjustment that would solve this problem. The reason for that is unlike for the QED or for the benchmark design, you cannot assume that changes over time would be the same in apple and orange trees. Not only are they different kinds of fruit, but they benefit from different climates. So a climate that helps apples might hurt oranges. So there's simply nothing you can do if you find a comparison group that's this bad.

So the key takeaway that I would like to emphasize from this series of examples is that the more effort you put into your design in terms of selecting a good comparison group, the easier, and the more convincing your analysis would be. So we'll spend the rest of the webinar talking about how to actually go about selecting a good comparison group. But before I do that, I would like to pause and take one or two questions if anyone has typed any in.

So Kim, do you have anything? I can't see the questions.

Kim Francis: Yes, there are a couple questions. I think both of them look like they might be for OAH actually. I'll just read them real quickly. Is it 5% or 10% of our work that needs to be evaluated? I've heard both percentages.

Amy Farb: Okay, this is Amy Farb. Five percent to 10% is actually related to observations for your performance measures. So that requirement is 5% of the program, 5% of what you're implementing needs to be observed. So that's not actually related to the impact evaluation. That's related to performance measures.

Kim Francis: Great. Great. And then we just have one more. When may we receive a copy of the PowerPoint and audio for this webinar?

Amy Farb: So we have to wait for the platform that we use to get it all back to us. It has to be made 508 compliant and then we'll be posting it for you. So I'd look for it over the next week.

Kim Francis: Great, thanks. And those are all the questions we have for now.

Randall Juras: Okay. Well, if you have any other questions as we're going along, please type them in. But for now, I'll move on. So now, I would like to draw your attention to another critical feature of the previous examples. You might have noticed in each of those examples, although the unit of measurement was the apple -- in other words, we measured the number of seeds per apple -- the unit of comparison for the purpose of selecting a comparison group was the tree. In other words, we selected comparison trees and then obtained the outcome by measuring the number of seeds per apple on those trees. Something we could have done that would not have been as convincing would be to select similar looking apples without regard to what kind of tree they were on.

So for example, we could have treated all of the apples on the bottom branches of one tree and compared them with a group of untreated apples on the bottom branches of another tree. However, if the treatment and comparison apples were on different kinds of trees, for example a Macintosh tree in the treatment group, and a Golden Delicious tree in the comparison group, this would not be a convincing comparison. And you may be thinking to yourself that sounds trivial. It's obvious, but this is directly analogous to an evaluation of community wide teen pregnancy prevention intervention. And it's something that in technical assistance we see quite a lot. So it's a point that I'll come back to and hopefully reinforce later in the presentation.

The basic idea is that instead of comparing similar people across two or more communities, it's much better to compare people across similar communities, which are defined using community level attributes. So use community level attributes to select into your comparison group communities that look like the communities where you're implementing your programs. And then compare individuals across those two communities, either with or without regard to what their individual attributes are.

So what all of these examples really have been illustrating, and that's one of them, is an instance of potential confounds. The key challenge in selecting a comparison group is to remove as many potential confounds as possible and the confounds or confounding factor is something I'm defining as a component that is completely aligned with one study condition, which would make it impossible to separate the effect of the intervention from the effect of the confound. Meaning that you could not attribute the impact to the intervention. For example, if all of the trees in your treatment group were Macintosh apples and all of the trees in the comparison group were Golden Delicious, you could not separate out the impact of the fertilizer from the impact of the tree type.

Likewise, if all your treatment communities are in the Northeast and all of your control communities are in the Southwest, you could not separate the effect of the intervention from the effect of geographic differences in socioeconomics or culture. So it's important that you choose a comparison group that overlaps in many ways with your treatment group on observable characteristics.

Common confounding factors that you might encounter include different data used to measure outcomes across groups. So for example, if you use surveys to measure outcomes in your treatment group and you use extent

administrative data to measure outcomes in your comparison group that would be a confound. Groups that are in dissimilar geographic locations is a confound, which is why you need to choose a local comparison group. And different demographic characteristics across populations is a confound, which is why you need to just use a focal comparison group.

So if, for example, all of the individuals in your treatment group are low income Hispanics and all of the individuals in your comparison group are high income Hispanics then income level is a confound and you won't be able to separate that out from the effect of the intervention that you're trying to measure.

So to summarize, the goal is to select a comparison group in which potential confounds are minimized. You can do this by selecting a comparison group that's local, focal, and which observations are aggregated at the same level and in which outcomes are measured using the same data source across groups. Once you've found the comparison group that meets these criteria, you should check to make sure that the two groups look similar at baseline. In other words, that the treatment group looks like the comparison group.

In other words, you should check to make sure that the treatment group and the comparison group look alike before the intervention starts. If they do not, the skeptical reader might worry that these preexisting differences are what's really explaining any post-intervention difference in the outcomes, rather than the intervention. To do this, to do a baseline equivalence test, you will need to have some characteristics that are measured prior to the intervention. The best of these, which is often available in extent data, is a baseline measure of your outcome of interest. If the outcome, for example, the teen pregnancy rate, is the same in both groups before the intervention starts, then any differences in

the teen pregnancy rate after the intervention ends can more credibly be attributed to the intervention.

Note that researchers increasingly prefer to report the magnitude rather than or in addition to the statistical significance of baseline differences and there are in fact guidelines published by various entities for how large of a difference measured as a standard effect size is acceptable. In fact, I believe that OAH has published some guidance on baseline equivalence for the first round of Tier B grantees on this topic and we'll make sure that that's made available to you. It's also publicly available on the internet along with the other evaluation technical assistance briefs that OAH put out for the first round.

This brings me to another topic, which is data or data sources. You can only answer a research question if you can measure the outcome and you can only assess baseline balance if you can measure the outcome both before and after the intervention starts. In order to find a well matched comparison group, you will probably also need data on things other than the outcomes, such as demographic characteristics, the unemployment rate in different communities, the racial mix in different communities, the average income level. There's no reason that the outcome measure and the variables used for matching or assessing baseline equivalence couldn't come from different data sources as long as each data source identifies communities in the same way so that you can combine the data together. In other words, those data sources have to be aggregated and identifiable at the same level.

For these evaluations, I believe OAH expects that you will use extent data, in other words use data sets that are already available because the data have been collected by someone else. In terms of the evaluation, it will not help you -- and let me repeat that -- it will not help you to survey youth in your treatment group because you will not have similarly collected survey data for the

comparison group. There are many, many potential sources of data that you could use that are already publicly available, or available by data use agreement with some agency, and we expect to have a future webinar on this topic. So I'm not going to go into detail here.

So now that I've talked about key characteristics of the comparison group in theory, I want to spend a little time talking about how to select a comparison group in practice. If you had a very small sample, a small number of observations and each observation having a small number of observable characteristics, it would be possible to select a comparison group by hand using, for example, exact matching. For example, if you have a state level intervention, there are only 50 states in the U.S. So you could potentially get that whole list of 50 states and eyeball them to see what's a good comparison. In a small sample like that, it's easy to keep track of all of the key factors by hand. You want to make sure essentially that the comparison group is local and focal.

So here's an example to illustrate what's going on. Suppose that you have a large number of treatment observations of which the observation for state A is one. They're on the left. State A is in the Northwest. It has a 5% rate of the outcome occurring, teen pregnancy or whatever it is. It has a 25% poverty rate. And you want to know what's the best comparison for state A, and you have a list of potential candidates from other states, the ZYXWV and U, which are shown on the right, along with their region outcome rate and their poverty rate.

So you want to select a comparison group that's local. So looking at your comparison, you see State A is in the Northwest while States W and U -- W and V -- are also in the Northwest. So they're local. You want a comparison group that's focal, has the same demographic characteristics. And here, you

can see that states X and W have essentially the same demographic characteristics as the treatment observation does. So combining those two things, you see that State X is both local and focal. So you think it might make a good comparison, and then of course you want to check for baseline equivalence of the two.

So your key outcome has a 5% prevalence rate in both samples. So you would conclude that this is a good match. It would be possible, of course, to select more than one comparison for State A, if you wanted to. In that case, it looks like none of these states are as good of candidates as State X, but you could also go to comparison states -- if you think they're appropriate enough -- that are either local or focal but not both. So you could go to State W, for example, which is focal and exhibits baseline balance, or State V, which is local but doesn't exhibit very good baseline balance on the outcome measure.

Of those, I would recommend selecting state W because it's balanced. Unfortunately, with a large number of variables or if you have a lot of observations, this can get out of hand quite quickly. Fortunately, you can and probably should let the computer do it for you. There are methods such as propensity score matching, PSM, and Mahalanobis matching, which people often call maha matching that has been developed specifically for this purpose and are widely implemented in all of the popular software packages. These things, they can get pretty intense sounding. If you look them up, there's a lot of math involved but essentially, they do the same thing as hands matching but in a rigorous way.

In other words, they answer the question for each community in the treatment group, which community or communities and the comparison group looks most similar, only it's defining most similar in some sort of rigorous mathematical way. I do want to emphasize strongly that software based

methods like propensity score matching work best when the pool of potential comparisons is already as similar as possible to the treatment group. That's because matching can only account for observed differences between the two groups and it cannot account for unobservable ones. And in fact, it can't account for all observed differences all at once always.

So it is much better, even if you're going to use software based matching, to try to find a comparison group that looks as similar to the treatment group as you can before you try to do the software based matching. Another thing to watch out for is that matching should be done at the level of selection to the treatment group. So if you're treating whole communities you want to match to other communities. And this is a mistake I see that's made fairly often using propensity score Mahalanobis matching. You do not want to pick individuals in other communities that look like the individuals in your comparison group, or at least not in the first step. In the first step, you want to select other communities that look like the communities that you're treating, and then as a second step, you could also try to narrow it down to individuals that look like those in your treatment communities. This is the same thing as comparing apple trees, instead of apples that look similar across two different types of trees.

Most statistical software will readily perform propensity score matching or Mahalanobis matching, although oftentimes it's not integrated into the core software. So you have to download a user written routine. But it is fairly easy to use. For example, using PS Match or PS Match 2 and (Stata). We expect the grantees that are chosen for the federal evaluation will implement some kind of software based matching, and we will help them intensively with that. We hope that many of the other grantees that are not part of the federal evaluation will also try some kind of matching and we do expect to provide

additional resources on how to do that, although probably not extensive one-on-one TA.

After matching, please remember that it is a good idea to check whether the match -- the groups -- after any statistical adjustments you've made to them -- are balanced on the baseline measure of the outcome. And with that, I would like to stop for a moment and take any other written questions that we've gotten.

Kim Francis: All right, we have some good questions. The first question is how do you define local?

Randall Juras: So local is difficult. It depends on your situation but what I think I would emphasize in this case is that you want to select comparison communities that are close enough to your treatment communities that you don't think that there would be huge cultural differences between the two groups of communities. We know the different parts of the U.S. have much different views on pregnancy prevention and much different pregnancy rates and trends in pregnancy. So I would try, if you can, if you're implementing in a single region of the country to select comparison groups that are basically within that same region.

And this, of course, is tied up with the idea of being focal. Local is almost a subset of being focal. You want your communities, basically, to look similar, your treatment and comparison communities. So if you have two communities, one a treatment community and one a potential comparison, and you're wondering if they're really close enough together than you could just take that second step and ask, well, do they look similar in enough ways that I'm convinced that there's not anything very different going on between the two groups or between the two communities.

So look at measures of poverty. Look at measures of the baseline measure of your outcome. Look at the racial characteristics. And if those things all look similar then it's probably a pretty good bet that you're going local enough.

Kim Francis: Okay. Another question, which might be a good segue into our next section, says, can we use the same group/local area over time? That is the same question comparing the year 2010 and every year going forward. I'm assuming they mean without a comparison community.

Randall Juras: Well, so I think the short answer is that without a comparison community that is very difficult to do. Not impossible, but difficult, and there are some slides in the next part of the presentation that specifically address that issue. So I'll leave that at that for now and I encourage you to ask again if I don't give you enough information in the next segment.

Kim Francis: Okay. Another person is asking, just to be very clear, you don't want us to collect any surveys directly from the youth. And I'll just say real quickly and then Randall, you can chime in. For - you may be collecting surveys directly from youths participating in your evidence-based programs for the purposes of participant satisfaction, program improvement, and feedback, and things like that. And which is up to you, but for the purposes of like the impact evaluation of the community-based intervention, it would not be advised.

Randall Juras: Right, please do not collect any survey data, which is specifically related to you're -- specifically for the purpose of your impact evaluation. But yes, you may need to collect survey data for other purposes.

Kim Francis: And I'll do one more quick question and we'll move on. When do we find out if we were chosen for the federal evaluation? So we are hoping and planning

to start reaching out to potential candidates probably in the month of May, what we're shooting for. And I think we'll better move on so we can get through the last section here.

Randall Juras: Okay, thanks Kim and if you have any questions, please type them in. We'll probably start with those after the end of the next section, but then open it up for verbal comments.

So continuing on, even with a well-matched comparison group that shows baseline equivalence, you should check to see if you can obtain data on the outcome of interest for several time points before the intervention began. And this is sort of getting at the last question, third to last question. Extent data sets often include historical data. In other words, if you're going to a state agency to get data on the teen pregnancy rates in communities, it's likely that they've been collecting that data for many years. And so you could get that data hopefully going fairly far back in time.

If so, you can sue those historical observations to considerably strengthen your analysis. There are several ways of doing this. The method that I want to highlight here to demonstrate why it's important is something that we call the comparative to short interrupted time series, which some people call CSITS, but I'm just going to call CITS. The basic idea behind CITS is it allows you to find a comparison group with similar pre-intervention trends rather than just similar baseline characteristics. Or more accurately, it allows you to use statistical adjustments to control for pre-intervention differences in the outcome trends.

So here is a very simplified example of how a comparative interrupted time series works. This figure shows time plotted on the X-axis with the intervention occurring right in the middle of the timeline, and a good outcome

on the Y-axis. In other words, something in which a better -- a higher number is better. I put on here the graduation rate. You can see that -- on the graph -- there are two post-program measures of the outcome, one for the treatment group and one for the comparison group. And the treatment group here appears to be doing much better than the comparison group.

Now, say that you are able to obtain data for several time points before the intervention started. You could plot these on the figure, as you see here, and, well, I meant to make these lines almost but not quite parallel, but on this slide, they actually look fairly close to parallel. Now, you could extend those trend lines into the future to see what you would expect to happen in each group in the absence of the intervention. As you can see, the comparison group observation happens to fall just about where we would expect it to, although this won't always happen, whereas the treatment group observation is a little bit better than we expected it to be, just extrapolating out that trend.

The impact that we measure is not the difference between the post-program treatment and comparison observations. Instead, it's the difference between what we expected to happen and what actually happens. Although comparative interrupted time series analyses can be fairly complex, this basic example I hope shows you in principle how such a method can improve the quality of your analysis. And this method would work to adjust for trends, pre-intervention trends between the treatment and comparison group, even if those trend lines are not parallel. But you do need to measure enough time points before the -- from before the intervention that you can establish a trend in both the treatment group and the comparison groups.

So now, having talked about what you should do, I would like to walk through a couple of examples of research designs that, while sometimes acceptable, can be very misleading in community level research. So if you are planning

one of these designs, I encourage you to proceed with caution. They are using an individual level comparison group, doing a pre, post, or interrupted time series design without a comparison group, or benchmarking to state or national trends.

So the first of these designs is using a within the community individual level comparison group. So the idea here is that you are - you have individuals in your community who are receiving the treatment. For example, they're participating in an evidence-based program or they are in a school that's participating in an evidence-based program. And you go and find other individuals in your community who look like the participants in your evidenced based programs and use them as a comparison group. Then you can measure the outcomes for your participants and for the comparison individuals, and see which group of individuals did better.

This looks a whole lot like what many of the Tier 1 grantees were encouraged to do in the last round. If you were a previous Tier 1B grantee, you may have done a randomized controlled trial using an individual level control group. The reason that worked for the last around is because there was an assumption made that the number of individuals who were treated in a community was very small relative to the total number of individuals in the community. So that the comparison group individuals, or control group individuals in that case, were likely not very much affected by the program.

In this context, the current round of grants, that assumption is problematic because many of these interventions are designed to saturate communities, or to be scaled up to as big of a scale as they possibly can be in the communities. So it is no longer appropriate to make the assumption that the individuals who are not directly receiving the EVPs are unaffected by the program. They very well may be because their boyfriend or girlfriend is participating in the

program, for example, or community norms are changing. So I would encourage you to stay away from this sort of design, if you possibly can.

The second potentially problematic design is an interrupted time series design. That's, as the name implies, a lot like the comparative interrupted time series design except that it uses a single group over time to measure the impact. So you would measure the trends for the treatment group only over time before the intervention begins and then see if your post-program observations differ from the trends that you established before the intervention started. The interrupted time series design is and can be a very rigorous design when you have a lot of observations so that you can establish a very clear trend in the outcome before the intervention.

Unfortunately, interrupted time series does require outcomes to be measured at many points over a long time span, and Shadish, Cook, and Campbell, in their book called Experimental and Quasi-Experimental Designs for Generalized Causal Inference, assert that at least 100 observations are required over time for this kind of design to be credible. It's very unlikely that many of you will be able to find 100 observations, in other words observations at 100 different time points for each of your treatment communities. And with only a handful of observations -- which is really anything significantly less than 100 -- this method, unfortunately, is simply not very convincing. And in the extreme case, if you measure just one time point before the intervention started and one time point after, it's what we call the much dreaded pre/post design. So if you are going to do this kind of a design, be careful before proposing it that you can get a lot of outcome measures over time.

The final potentially problematic design that I want to talk about is benchmarking to state or national data. In a design like this, you would measure outcomes for your treatment group and compare them to state or

national averages of that outcome. This is a research design that if done correctly can sometimes yield a convincing estimate of the program impact. Reading through the applications that you guys put in, we have seen that several of you proposed some variation on this kind of design. And we don't want to tell you here that you shouldn't do it, but we do want to warn you to proceed with extreme caution if you were doing it.

The reason that this design could be problematic is that the trends in your carefully selected treatment communities may be much, much different than state or national trends. One thing that we often see is that program implementers target communities that are much different than the average on purpose. But these communities, which are basically outliers, tend to revert to the average over time even if no programs are implemented. So we know, for example, that the pregnancy rate over time nationwide is slowly declining. In communities that have very high teen pregnancy rates, those pregnancy rates could be declining much faster than the national average because they are becoming more like the national average.

For that reason, it's usually better to benchmark to national trends rather than to national averages at a single point in time by applying a design like the difference in differences design or a comparative short interrupted time series. Just to reiterate, you could be seeing something like this where you have, again, time on the X-axis and this good outcome -- the graduation rate in this case -- on the Y-axis. You see in your comparison group, which is the national average that the intervention is -- the outcome is slowly improving over time, even without the intervention. The bottom trend line shows that the treatment group is becoming more like the comparison group over time, even before the intervention is implemented. So that if you were to only observe the treatment and comparison groups after the intervention was implemented -- that's those big dots on the right hand side -- you would come to the conclusion that the

program was very successful. So that would be if you were benchmarking to national trends but without any data from before the intervention started. You would conclude that this program was successful.

However, in reality, the post-program observations are, as you can see from the extended trend lines, just exactly what you would expect to see without the intervention, if the intervention were not implemented. And I don't want to mislead you into thinking that this will always make your program look good. It could also make your program look bad for some outcomes. This kind of thing is very common in benchmark designs. So if you are doing a benchmark design, please be very careful about it.

So that's all I have for you in today's webinar. To summarize what I've talked about, selecting a comparison group begins with a well-specified research question that includes the intervention name, the target population, the counterfactual condition, and the outcome domain. Once you have the research question well specified, you should try to select a comparison group that looks like the counterfactual you specified in your research question and avoid compounds. You can try to avoid compounds by thinking about keeping it local and focal and measuring for units that are at the same level of aggregation and using the same data source.

Then after you've tried your best to select a comparison group on those principles, you should match the treatment and comparison groups, possibly by using software based methods like propensity score matching or Mahalanobis matching. And then once you think you have the perfect comparison group, test for baseline equivalence. Look at the outcomes and other demographic variables from before the intervention started and see if they look similar in the treatment and comparison groups.

Finally, apply any statistical adjustments, if necessary, which is what we call analysis, using something like a difference indifference design, a comparative interrupted time series design, or if it's appropriate, just a simple linear regression model. So now, I'd like to open it up for any other questions.

Kim Francis: Okay, I have a few more that have been typed in along the way. So maybe we can start with those and then if folks would like to ask a question directly, we can do that. So one question is what is the minimum number of potential comparison communities that using statistical matching strategies is possible? What's the minimum number of potential comparison communities in which you can use statistical matching strategies?

Randall Juras: That's a very good question and I think the technical answer is probably two. But the practical answer is many more than that. So technically, you can only match -- you have to have a number of observations, which is larger than the number of characteristics on which you're matching. If you have only two observations in the comparison group, you can't match on three different variables -- the outcome of interest, and the racial composition in the community, and the economic climate, for example. You would need at least four.

So the strength of your matching will increase as you add more and more communities in the comparison group. I think that the easiest answer here, or maybe the most practical advice I could give you is to try to select as many comparison communities as you possibly can without straying too far from the ideal that they should be local and focal. But other than that, there's no solid answer, I think, for how many communities you need just for the method to work, technically. How many you need to arrive at a statically significant answer, after you've done the impact analysis, is a little bit different. There, the answer is that you would need to perform a statistical power analysis to be

able to know whether you would expect to have enough power using the sample size that you have to arrive at a statistically significant answer. And that involves knowing some things like the expected variants and the outcome measure, which you can get from other data. And the amount of variants -- the amount of variability within the treatment and comparison groups that's explained by different characteristics.

If you want to start getting an idea of what your power would look like, or how to do a statistical power analysis, there is a resource on OAH's website that was part of the first round of technical assistance for the first round grantees on how to do a power analysis. There's also a really wonderful paper that begins to explain it, by Howard Bloom at MDRC, which is called the core analytics of randomized experiments for social research, I believe. And it starts to walk through how to do a power analysis, how to do a basic power analysis, and how the power analysis changes if you're doing a community level design -- which he calls a cluster randomized trial -- rather than doing an individual level design.

There's some software that will perform a power calculation for you fairly easily. There's one called Power Up, which I like quite a lot. There's another one called Optimal Design that a lot of people use. So you should go to one of those kinds of software to start doing a power analysis if you want to know how big of a sample you need to be able to detect impacts of a reasonable size.

Kim Francis: Okay, great. We have more online questions so I'll just keep reading those. We have nine communities as treatment communities. How many of them should be comparison?

Randall Juras: I'm not sure I completely understand that question. If you have nine treatment communities than they're all treatment communities. You should try to find at least one comparison community for each one of those treatment communities, but I strongly suspect that a power analysis would show that one comparison community for each of the treatment communities, or for a total sample size of 18, is probably a little bit small. So again, you know, I would go out and try to find as many potential comparison communities as you can that basically pass the sniff test and then use some kind of statistical matching procedure like propensity score matching to rule out any that really turn out not to be appropriate. And that is one of the nice things that propensity score matching routines do.

And in fact, I believe one of the reasons the propensity score matching was originally developed was to try to determine whether some of the comparison group was simply too dissimilar that it wasn't appropriate to compare it with the treatment group. And so by looking at the amount of overlap in the propensity score between the treatment and comparison group, you can knock off potential comparison communities so that you end up with a good set. But again, just I would try to go out and find as many as you can that pass the smell test and then use those procedures to narrow it down if need be.

Kim Francis: Great. Here's another good question. Our two communities are military communities. Should we be looking at other military communities for comparison or should we just focus on matching on the outcome measures, such as birth rates?

Randall Juras: Well, in that case, you know, it depends on how much you want the reader to believe your results. And I think that the answer there is that if you have -- if your community has some very strong characteristic like being a military community that you think makes it much different than the surrounding

communities then having a comparison group that's focal is much more important than having a comparison group that's local.

So for a military community, I would advise trying to find other military communities rather than trying to find local comparison groups that look similar based on the baseline measure of the outcome. Because again, even though the baseline measure of the outcome may be the same, other trends may be much, much different in non-military versus military communities.

Kim Francis: Great. How many characteristics are needed to identify a good comparison? This is kind of a two-part question but we'll start with that one.

Randall Juras: Okay. Well, to not answer the first part, I don't know that there's a single answer to that question. Again, what you're trying to do here is to convince a skeptical reader that you have a comparison group that looks like the counterfactual that looks like the treatment group would have looked if it had not received the intervention. And the more evidence you can bring to bear to show the reader that that's true, the better.

So if you're only able to find a dataset that has your outcome of interest and one other characteristic, you can only use that one other characteristic and maybe intuition to find appropriate comparisons. If you have many more characteristics than that, you should choose the best one. There is a maximum. You don't -- I think that the current best advice is not to throw the kitchen sink at the problem, but to try to go through in a principled way and find a set of variables that you think would be correlated with your outcome of interest and possibly explain selection into the treatment group, so that you can arrive at the most convincing comparison possible.

But there's really no minimum number and there's no maximum number. It's just that you're kind of like a lawyer trying to make a case. You have to demonstrate to the reader that the comparison is a good one and not a bad one. So use as much evidence as you have to make that comparison.

Kim Francis: Okay. And the second part of this question is, in our area, we have four communities. We are likely to not find comparison counties within our state. Is taking a state level approach more appropriate than finding counties out of state or region?

Randall Juras: I think that that's a balancing act. Unfortunately, there aren't black and white answers to a lot of these questions. You have to decide in a case like that, you know, what's the appropriate balance between focal and local, realizing that you're not seeing everything about what's making a local comparison group a good comparison or a focal comparison group. So, you know, perhaps a faraway comparison group that has some of the same measurable characteristics is not believable, but maybe a local comparison group, with much different observable characteristics, is also not very convincing.

So you may need to strike a middle ground, going to, for example, neighboring states, or nearby states, or states in a region of the country that you think is kind of similar to your region of the country and looking for a comparison group that exhibits some of the same focal characteristics -- in other words, similar demographics. But it's really a balancing act. For example, if I were doing an evaluation of a teen pregnancy program in Massachusetts where I live, in communities here, I would be much more comfortable comparing those communities to communities probably in places like -- even if I had to go out of region to places like, you know, Washington State or California then I would be comparing it to communities in Oklahoma or Texas, simply because there are many different -- there are a lot of cultural

and other socioeconomic differences between the states that even if they don't show up in your measurable set of characteristics, you know that they exist and your readers will know that they exist.

So, you know, the short answer is you're going to have to try balance what seems right for finding something that's as local and as focal as you can.

Kim Francis: Okay. This question asks, is this a four-year impact evaluation? And I think the answer to this depends on what kind of data you're using and what the time lag is in receiving those data to answer sort of how many years will this cover.

Randall Juras: Yes, this question came up in the orientation meeting as well. Some of you might remember and I thought it was a good question then. I still think it's a good question. You know, you have to find data with which to measure the outcomes that are part of your research question. You're going to probably have to use extent data, which is collected by other people, and unfortunately sometimes that data is not made available to researchers for some time after it's collected. And it varies greatly by data source. Some data are available almost immediately and other data lag by years or more.

And so part of the challenge for you, I think, is to find an acceptable data source. And again, we'll have another webinar on this later. But you should find the data source that balances the things you're looking for, that measures the outcome in a reasonable, credible way that has face validity. Hopefully, the measure is some other characteristics that you can use for matching and it's available to you pretty quickly. I realize that these grants ends not so long after the program has started being implemented and you need time for reporting.

So you may not be able to measure outcomes -- depending on your data source -- for very long after the program starts. On the other hand, it's not entirely clear with programs like this yet, as far as I know, it's not clear when you would expect to see the peak impact. Is it just after the program ends? Is it during the program? Is it a year after the program ends, or five years after the program ends. We'd love to know all of those things but unfortunately, you won't be able to go very far out, I think, with these impact evaluations. You'll have to look at some short-term outcomes probably -- I couldn't give you a number of years. I'm not that familiar with the timeline. Perhaps Amy could. But outcomes that are measured not very long after the intervention starts. Did you have any thoughts on that, Amy?

Amy Farb: No, I think you -- I was jumping in just as you jumped in, Randall. I think you handled that and we can prepare a nice written response to that as well that we can share with everybody after the call.

Randall Juras: Okay. Thanks.

Kim Francis: Are there any questions over the phone?

Randall Juras: Is everyone unmuted?

Coordinator: To ask a question on the phone, they can press the star 1 and please record your name. Again, press star 1 to ask a question. One moment. We do have a few questions coming up.

Kim Francis: Okay, great.

Coordinator: The first question comes from (Nicole Durskey). Go ahead.

(Nicole Durskey): Hi. Hello. Thank you. Our community is an entire county. So I'm just trying to think about what the unit would be for a comparison community. Would it be the county level or I'm just trying to feel out how we might approach that.

Randall Juras: Yes, I think the short answer to that is that if you have a program which is being scaled up to an entire county and you expect to serve, you know, participants who are distributed throughout the county, then the most appropriate comparison group would be a set of other counties that resemble the county in which you're delivering the intervention. Hopefully, you know, nearby counties, if you could find them. But of course, the number of counties in any given area is limited. So you might have to go a little farther. That's a challenge with community level designs, with cost of designs like this.

But yes, it's hard to see how any other comparison group level would work very easily. So yes, it seems the counties is probably the right answer.

(Nicole Durskey): Okay, thank you very much.

Coordinator: And we have a question from (April).

(April): Hi, yes. Our question was if the valuation -- the impact to the valuation start with the implementation of year two and is it expected that every year we are continuing to evaluate that impact?

Amy Farb: So this is Amy Farb. That's not necessarily our expectation, although that would be a perfectly acceptable strategy. If you wanted to go, you know, from year two and you had data, you know, two years later, three years later, that's perfectly fine as well. Does that answer your question?

(April): Yes. So does that -- well, I have one question then following would be then does impact evaluation have to start in year two or is there, I guess, in year two or could it even possibly be impact evaluation on year three possibly?

Amy Farb: I think it should begin by the time you're trying to implement your project because you're starting to change things already and you don't want to try to measure a baseline or get baseline data when you've already started implementing your project.

Randall Juras: Yes, so another way I think maybe to answer that, building on what Amy said, is that you're measuring the impact of the program. If you're doing that in a way that involves data over time, which we hope you can, you would want to measure the level of the outcome before your intervention began, just so that you can make sure your treatment and comparison groups are comparable. Then you want to measure the impact again after the program starts. And you could measure it during year one and then during year two, and then during year three.

But a lot of the practical work of doing the impact evaluation would probably come towards the end. It's not like you need -- from an evaluation perspective, and I don't know about from a grant management perspective, but from an evaluation perspective, you wouldn't necessarily need to be trying to track that data all along. You could gather it at the end and then go back and see what the impact was after one year, after two years, after three years.

Amy Farb: Thank you.

Coordinator: And we do have a question from (Mary Langley).

(Mary Langley): Okay, I just want to be clear about the impact evaluation because we are doing baseline data with community readiness. Because if you're only serving like 250, even in a small rural county, you still will need to environmental strategy to measure the impact on the community, the teenage pregnancy effort, not just the participants in the program. So is that a measure where were going to conduct the community readiness at the beginning and then later on to see whether or not the community is aware of the effort. Because sometimes the data about teen pregnancy live birth data is a couple years lag. So you wouldn't have -- even in comparison, you wouldn't have the data by the time the grant ends.

So I'm just -- I'm not the evaluator so I'm just asking a lay person -- I'm asking a project (unintelligible). So but some clarification.

Randall Juras: Yes, I think that the -- if I'm understanding what the community readiness measure is, that actually sounds to me like it would be more a part of the implementation study than the impact study.

(Mary Langley): Okay.

Randall Juras: To see what, you know, you're measuring it only in your community that you're going to and the purpose, and tell me if I'm wrong, is to see whether the community is aware of what you're doing. That's not what we think of as -- that's not an outcome that you usually measure in an impact evaluation, although it is an important measure. It's important information to have to inform the results of the impact evaluation. It's not itself an impact on one of the things that the program is designed to change, like pregnancy rates or teen birth rates. But you should measure it. You should collect that data and you probably want to report on it or could report on it earlier than the results of your impact evaluation.

And yes, unfortunately, it sometimes measures for, you know, your outcomes of interest, things like teen pregnancy rates or teen birth rates lag, which will make it challenging for your evaluator to complete the evaluation within the period of the grant. But that's -- I don't know what to say except that that's a challenge that we have and so, you know, they should try to seek out data, which is available as quickly as possible to answer questions as quickly as they can.

(Mary Langley): Yes, I think that the community (unintelligible) more, like you say, not impact but process implementation. But it's also sustainability because as the numbers increase, the community is more vested in sustaining those efforts. So I guess it could be added to the quality of the sustainability plan. Okay. Thank you.

Coordinator: All right, we have one more question. One moment. Okay, we have a question from (Luanne Roebuck).

(Luanne Roebuck): Hi. It sounds like primarily what you and OAH are looking for as far as an outcome measure for the outcome evaluation is teen birth rates. Am I understanding that correctly?

Randall Juras: There are a lot of outcomes that I would think would be appropriate and I'm going to let Amy jump in for -- after I'm done -- to tell you the -- what OAH is actually looking for. But it seems to me that any outcome which is really strongly tied to your logic model would be a good outcome.

So you're programs could be changing. They could be changing sexual risk behavior, for example, and then, you know, some logic models would specify that as a result of changing sexual risk behavior, they change the pregnancy rate and then by changing the pregnancy rate, they change the birthrate.

Because the birthrate is sort of the ultimate goal, it seems like the best outcome, in a sense, to measure. Unfortunately, birthrates probably lag quite a lot behind other things that you can observe more readily that are sort of more proximal to the intervention, like sexual risk behavior.

Along with that might be things like school attendance or other academic outcomes, like graduation or college attendance, or something like that, that you might be able to measure. If they're really tied to your logic model, might be able to measure easily and have available to you much, much quicker than things like the teen birthrate. So although teen birthrate, you know, probably would be a very convincing outcome to measure at the end of these evaluations, you know, I am skeptical that many of you would be able to do that because of the short timeframe of the evaluations.

So you may want to focus on outcomes that are more proximal to the intervention, like, you know, sexual risk measures, although those might be hard to get from extent data, or at least something like pregnancy rates, which might be visible a little bit quicker than birthrates.

And Amy, did you want to add anything to that or contradict me?

Amy Farb: No, that was beautifully answered. I mean, ideally we would love to see impacts on birthrates, but by the end of your grants those will probably still be two, three years behind and you might not be able to detect any impact there. Although I would definitely look at it and see if you can, but we're flexible. So as Randall said, anything related to the logic model. That's what we're interested in.

(Luanne Roebuck): Right. So just a follow-up comment. It seems like sexual risk behaviors are a logical, more proximal outcome. But as you pointed out, Randall, those

are often challenging to -- data to get, particularly aggregated at a community level in a way that would make sense as a comparison community to your target area. So if we could address that in the next webinar that would be great.

Randall Juras: We'll try.

(Luanne Roebuck): Okay.

Coordinator: All right. And we do have a question from Mary Langley.

(Mary Langley): No, I answered mine. I mean I asked…

Coordinator: Okay, all right. All right, and there's no questions at this time on the phone.

Kim Francis: Okay, and it looks like we're out of time. And I know there was some additional questions typed into the little chat box that we don't have time to address right now, but we will compile these into a written document with answers to them and make sure that you have access to those. And in the meantime, if you have specific questions that come up about your evaluation design, please continue to direct those to your project officers and we'll work on getting you a response. And we'll be, as we said before, this webinar recording and transcript will be made available to you shortly as well.

And Amy, do you have any closing words?

Amy Farb: Thanks everybody for joining us this afternoon. I think this was a really important webinar and I am so pleased that most of you were able to get on and hear it today, and we will continue to follow-up. And I hope everyone has a wonderful afternoon.

Coordinator: All right, thank you. This completes today's conference. You may disconnect at this time.

END