[Tara Rice] Hello everyone and welcome to our webinar. On behalf of the Office of Population Affairs, I would like to welcome you to this webinar on Summative Evaluation Approaches. This is one of a series of webinars on evaluation that's being delivered by Mathematica, OPA's evaluation technical assistance provider. These webinars are open to anyone who's interested in learning more about the evaluation approaches used within various types of OPA funded projects. I'd like to just set the stage for what you can and cannot expect on this call today. We will be discussing summative evaluation approaches that may be useful for conducting rigorous and useful tests of program affects. We will discuss the conditions under which summative evaluations are appropriate and provide general recommendations on how to design and implement a credible test of program affects. We will also be answering general questions about the content presented on the webinar. You can enter questions at any time in the chat. And we'll also offer time at the end of the call for our presenter to answer questions. But on today's webinar, we will not be answering questions specifically about any of the TPP NOFOs or other grant programs and we will not be answering questions about individual proposals or provide specific guidance on an individual grant application. If you have NOFO specific questions, those should be directed to the contacts that are listed within the NOFO. So at this time I'd like to turn over the presentation to our main presenter, Lauren Scher from Mathematica. Lauren, the floor is yours.

[Lauren Scher] Thank you Tara. Hi everyone, I'm Lauren Scher and I'm a senior researcher at Mathematica and have been providing evaluation technical assistance to teen pregnancy prevention grantees funded by OPA for over a decade. Currently I'm a project director for an effort to provide evaluation technical assistance support for grantees who were funded in 2020, including grantees who are implementing rigorous impact designs. In addition, I previously designed and implemented impact evaluations and I currently direct a project funded by the U.S. Department of Education, to help researchers design and implement studies that will support evidence use. So, today's presentation builds on earlier presentations in this webinar series, by now focusing on summative evaluations. In particular today, I'll discuss designing effectiveness evaluations, also known as impact evaluations, we'll talk about the circumstances under which impact evaluations may be warranted, and then walk through various design, analytic, and reporting considerations. In particular today, we're going to focus our time on understanding when and how to design an effectiveness study of an intervention.

In the first webinar of the series, my colleague, Gene Nabb, talked about that there are lots of evaluation activities that you may want to engage in before jumping into an effectiveness study. In fact, there's a real danger in jumping in too soon to examine intervention impacts if you haven't yet established that a program can be implemented with fidelity and shows early evidence of promise. But once an intervention is ready for testing, a well-designed and well-implemented summative impact evaluation holds promise in helping program providers, participants, community members, policy makers, and the broader TPP field, to feel confident that they understand whether an intervention has achieved its intended goals for the population that it serves. Typically, this evaluation work follows early formative evaluation work that my colleagues discussed in previous recent webinars. Notably, as my colleagues have mentioned, an impact evaluation tends to occur downstream in the continuum of evidence after earlier formative evaluation activities, like needs assessments and early implementation and outcome studies. But today, we're going to focus largely on the next step, conducting an impact evaluation. And before we jump into this, I want to point out that impact studies do not need to be conducted in isolation. As I'll discuss a bit later, strong impact studies also benefit from additional evaluation work, for example, including a companion implementation study, which I'll talk about a bit more later.

An implementation study can help understand whether and to what extent the intervention was implemented with sufficient fidelity and quality and met the needs of the community in the context of the impact evaluation. And in addition, economic studies and systems change evaluation methods may be warranted in addition to an impact study to examine the cost implications for implementing an intervention and the complex systems within which an intervention is implemented. These other evaluation strategies won't be our focus today, my main goal in pointing this out is to drive home the idea that you should carefully consider what you hope to learn from conducting an evaluation and plan to incorporate research methods and data collection activities that will help you get the answers that you want and need to move forward. So, let's start by talking about why you might want to test the effectiveness of an intervention through an impact study. Importantly, you want to move in this direction if you're interested in generating evidence that the intervention is causally responsible for improving outcomes, which is an improvement over the kinds of information that you might have gotten from an outcome study which was discussed in the early webinars. You may be in a situation where you've developed an intervention, but it hasn't yet been tested rigorously, or it could be that the intervention was previously tested rigorously but has been substantially adapted. For example, if the content has been updated or the program providers are using different modes of delivery such as incorporating technology elements.

Finally, you may want to engage in an impact study if an intervention was previously tested rigorously but with a very different population or in a very different setting. In this case, the program may or may not have been adapted. To summarize, if you're in one of those situations, you may feel that conducting a rigorous impact evaluation and generating evidence of effectiveness is the next logical step for your intervention. As I just mentioned, you don't want to jump into a rigorous impact evaluation too early. If you can go down this list and feel like you've checked off all the boxes, then you may want to consider planning for a rigorous impact study. Impact evaluations are high stakes, resource intensive efforts and can yield unsatisfying results if you proceed to them before your program and your partners are ready for it. Again, formative evaluation, early summative evaluation work, and ongoing communication with partners and community members can get you where you need to be. The items you'd ideally want to check off this list before diving in include an environmental scan that suggests that the community needs the intervention and is interested in implementing it, evidence that previous participants and/or caregivers reported being satisfied with the program, previous evidence that interested parties found it compelling and a good fit, documentation that the intervention is well-defined and has been implemented with fidelity and quality. And ideally, there's also evidence that suggests that near-term outcomes improved. For example, changes in sexual knowledge or behavioral intentions. And that there's a well thought out logic model that suggests that longer term behavioral outcomes will also improve.

And finally, that potential program partners feel invested in the evaluation's goals and understand and are willing to comply with the methodological requirements of a rigorous test of the intervention. Without each of these elements in place, an impact evaluation runs the risk of yielding unsatisfactory evidence. For example, an impact study may run the risk of yielding results that suggest minimal to no impact when really, the issue was lack of readiness to implement or a lack of a clear contrast between the treatment and control groups. And while this checklist is a necessary condition for showing that an intervention is ready for an evaluation, it's not sufficient. There's a long list of considerations when designing and conducting an impact evaluation to set your program up for an informative and useful test of its effectiveness.

So, let's now assume that you've checked off all those items in that previous list, and you're ready to design a rigorous impact evaluation. We'll now talk about some design issues that you should consider to ensure that you plan the strongest test of the intervention possible. First, we suggest thinking about the appropriate study design to examine the impact of your intervention. Note that today we're going to focus on group design studies, including randomized control trials, RCTs, and quasi-experimental designs, or QEDs. I'll refer to both of these as group design studies as both studies, study designs, include an intervention, aka a treatment, and a comparison, aka control group. In the first webinar in the series, Gene Nabb highlighted the importance of having a comparison group, noting that through examining the differences in outcomes between equivalent groups, we can have more confidence that changes in outcomes over time are due to the intervention as opposed to some other explanation that we would if we simply looked at outcomes from one group both before and after an intervention was implemented. Of course, study quality also matters in our ability to feel confident about the study findings, which we'll discuss shortly. But we also want to acknowledge that these study designs may not be appropriate in all situations. There are other study designs to consider, depending on the type of intervention you're testing and how and to whom the intervention may roll out. For example, a change to a broad policy may require a different kind of design entirely, such as a systems change evaluation.

In addition to RCTs and QEDs, there are also other designs that measure the intervention impact, such as interrupted time series, regression discontinuity, step wedge designs, and single case designs, which might be appropriate to consider. Which design you select depends on a number of issues, including the level at which the intervention is implemented, for example, testing a policy versus an individually provided curriculum, and the willingness of partners to engage in certain design requirements, like random assignment. Or, data collection requirements, like whether you can collect data at the individual or group level. So, now I have another list for you to consider, here's the list of elements that you may want to address as you plan group design impact evaluations. And for the reminder of today's presentation, I'll briefly discuss most of these issues. And at the end of the webinar, I'll point you in the direction of links to resources that offer a lot more specific information.

So, let's start by talking about research questions. When articulating research questions for an impact evaluation, your main goal should be focused on estimating the impact or effect of your program by comparing outcomes across the treatment and comparison conditions. The critical first step is making sure that you clearly specify research questions focusing on outcomes that align with a well-reasoned logic model. This includes specifying outcomes that you think are likely to change as the result of participating in the intervention within the time frame that you're able to carry out the study. It also includes thinking about whether you want to explore whether outcomes may vary for particular groups, for example, whether you think that impacts may be different for different age groups, genders, race, or ethnicities, or whether impacts may vary, for example, when delivered in different settings. Ideally, you'll have an idea about the types of outcomes appropriate for this impact evaluation based on early outcome study that you might have conducted. We encourage you to be realistic about what you can test within a specified time period and think about the kinds of outcomes that you may reasonably expect to change during that period of time. Even if your ultimate goal is to measure change and for example, long-term distal outcomes such as sexual behaviors and its consequences, you may want to include more near-term outcomes to examine earlier in the study, for example, changes in sexual behavioral intentions, beliefs, or precursor behaviors, which may help you understand whether your logic model or your theory of change is working as you would have expected.

And here we provide an example of a well-specified research question that mentions the intervention, which is, we're calling INOV, the outcome, which is sexual initiation, the time frame, which is 6 months after the program ends, and the population, females, ages 12-15, and the setting, middle school. So, when planning a group design study, you'll need to consider whether to implement an RCT or a QED. RCTs use a random mechanism to assign sample to two or more groups, in a well-implemented RCT we can feel more confident that differences between the groups on outcomes are due to the intervention. Randomization can occur at the individual level, such as the youth level, or at the group or cluster level, such as schools, teachers, or classrooms. Because of randomization, groups should be similar on both observed and unobserved characteristics. While an RCT may be preferable to a QED, it may not always be possible or feasible.

As you investigate design options, we suggest you consider asking the following questions to determine whether randomization to intervention and comparison conditions is feasible; first, consider whether there will be fewer slots than there are eligible participants for an intervention. In that scenario, randomization is an ethical and ideal way to select participants for limited slots. Next consider the appropriate or realistic level of assignment to treatment. Think about the level at which the intervention's delivered, for example, whether the intervention is delivered on an individual basis to intact classrooms or across entire schools, clinics, or community-based organizations. Even if an intervention is delivered at a group level, could participants be individually assigned to those groups? In school classroom delivered interventions, often whole classrooms may need to be randomly assigned to treatment and comparison conditions. However, in the case where you are signing up for an intervention, for example, an after-school program or a clinic-based program, then it may be possible to randomly assign at the individual level, even if the program may be delivered in a group setting.

Third, and importantly, critical conversations need to happen to determine whether randomization is feasible and at what level randomization makes sense. Close communication and collaboration can help ensure that the study partners feel heard, understand the importance of the study design, and are assured that the participants will be treated ethically. In fact, beyond determining study design, no impact evaluation will succeed in its goals without strong communication and collaboration with partners. Oh, I think I jumped the gun. Let me go back a little. And if randomization is not feasible, then QEDs are a viable option. QEDs use a less rigorous, non-random mechanism to form treatment and comparison groups. Even if groups appear similar on observed characteristics, such as age or race, they may be different in some unobserved ways. For example, those motivated to volunteer to participate in a program versus those who are not interested, thus differences in outcomes between groups may wind up being unrelated to the intervention, and for this reason, care what be taken to ensure that these groups are similar. And so, in a situation where random assignment is not feasible, you should think about how to best identify a comparable group. Here are some examples of QED methods that can help ensure similarities between an intervention and comparison groups. These include identifying an appropriate group to serve as the comparison or counterfactual, such as schools, clinics, community-based organizations, or whole communities. Based on characteristics such as type of school or size, school size, race and ethnicity, or financial hardship rates, like information on percent eligible for free or reduced price lunch, and ideally you'd have data on baseline measures of the outcomes of interest for your study but oftentimes that's not feasible so, you rely on proxies like demographics or other available administrative risk characteristics with the hopes that achieving balance on these proxies will yield balance on the baseline measures of outcomes.

Another way to try to ensure that you're comparing similar groups and could be used in conjunction with the approaches I just mentioned above, early, is to use statistical methods, such as propensity score matching or weighting to mitigate differences between participants. These approaches either pair back the sample to the subset of individuals who are well-matched in terms of demographics and baseline measures of outcomes or reweight the data to statistically make the treatment and comparison groups appear more similar than they would without weighting.

Finally, I'd like to call attention to a few methods that likely will yield comparison groups that will not be helpful to you in getting a credible or informative answer to whether your intervention is achieving its intended affects. These include benchmarking to national, state, or county level data. When you compare outcomes for your program group to national or state or county averages, you're likely going to miss attribute any differences to program affects when it's really likely that the underlying difference in the composition of your group and your benchmark sample are contributing to those differences. Another concern that you should be thinking about is solely focusing on the comparability of the comparison group without thinking carefully about the kinds of services that are available. A common concern when identifying a comparison group is whether that group is receiving similar services or might have the opportunity to receive the same services as the intervention group. This might happen, for example, if you throw in a neighboring comparison community and they might have easy access to the intervention that's being delivered one town over. In this kind of situation, the impact might be credible but the contrast will be very weak and it's going to produce an unsatisfying test of the effectiveness of the program.

Regardless of the study design you choose, there are a lot of considerations to think about at the design phase to ensure that you’re planning for a strong, useful impact study. So here are four important design considerations that you should be thinking about; lack of equivalence is a concern for randomized control trials, particularly those with high sample attrition as well as quasi-experimental designs. So it's really important to design a study to be able to measure and demonstrate the treatment and comparison groups included in your ultimate analysis were initially equivalent at baseline on observable characteristics. For this reason, a study should be designed from the beginning to specify characteristics to demonstrate equivalence, including age, race, ethnicity, gender, and baseline measures of the outcomes of interest. The Teen Pregnancy Prevention Evidence Review, or TPPER, outlines baseline equivalence requirements necessary to meet the evidence standards. Another important concern to think about at the design phase is determining the appropriate contrast to test. In an ideal situation, in terms of observing the largest impacts, we compare our intervention against nothing. This would be a very strong contrast in service experiences.

However, comparison groups either by design or by default often receive some sort of treatment, typically the business as usual, but if that business as usual overlaps with many key features of the intervention that you're testing, then you're not going to provide a satisfactory test of your intervention. So you're going to need to get in front of this, you need to design a study where you can expect the intervention setting will be substantially different from the business as usual condition. And you need to document the services that are available as part of the business as usual condition and how your intervention is different from it and plan to measure, to continue to measure this contrast in services during the life of the study. That's the best way to make sure that you're testing a sufficiently strong service contrast that will yield an informative test of the effect of your program.

You should also be careful to avoid confounding factors at the design phase. These are observed factors that completely align with one of the study conditions. For example, having one treatment and one comparison school is the most common confounding factor. We can't disentangle the scenario anything about the intervention from the single unit being used as the treatment site or analogously about the comparison group site. Another confounding factor is for example, when you use different data collection methods, for example, a paper based survey for intervention group members and web based surveys for the comparison group members. And then that might happen where folks might tend to respond more honestly to one type of survey method versus another. You don't want a situation where differences between the groups might be attributable to this confounding factor.

Finally, I'd like to talk a bit about determining an appropriate sample size. So regarding sample size, your study design team should use information that you may have learned from pilot studies to develop assumptions about expected recruitment, enrollment, and sample attrition expectations, as well as expected impacts. Using this information, you can get a ballpark estimate of the number of youth and clusters that you can serve during the window of time that you'll begin conducting the evaluation. This includes taking into account things like frontend planning period, and backend data collection, analysis and reporting. We strongly recommend that you be conservative in your statistical power calculations so that you can ensure that you set your impact study up for success. This includes being realistic about expected sample attrition rates. A particular concern, if you're working with hard-to-reach populations, such as those in foster care or involved in the juvenile justice system, expected attrition rates also have implications for the length of follow-up that you might reasonably expect for certain populations.

When conducting a statistical power analysis, you're going to need to determine the sample size large enough to detect statistically significant differences in your primary outcomes of interest. To do this, you should factor in the clustered nature of the design, if that's applicable, the strength of the contrast between the treatment and control groups, which we've talked about before, earlier, the minimum detectable affects size you expect to observe, given the expected sample size, and you can look into existing literature showing effectiveness of comparable programs to assess whether your study is in the right ballpark. And there's also tools and resources that you can use to conduct power analyses, and we provide links to resources at the end of this presentation.

So that's just a very, very quick discussion about statistical power analyses. We could go on for another 15 minutes or much longer than that, so we encourage you to look at the resources for more information on this topic.

So, beyond study design considerations, there's also important planning activities that need to occur at the outset. Importantly the design team should consider strategies for obtaining consent or assent. For an impact evaluation, you will likely need consent from participants for data collection efforts. This goes beyond the consent or assent for actually receiving programming. This includes thinking about the data that you're going to need and determining who will collect these data. You'll need to consider potential institutional review board requirements, or IRB requirements, and implications for your timeline as well as your study partner's comfort level with consent procedures. You're also going to need to develop clear protocols for obtaining connect information, conducting outreach, and providing incentives, because it's likely that you're going to need incentive, to incentivize data collection to get high response rates needed for credible estimates of program effectiveness.

In addition to data collection procedures, the procedures should be carefully thought through in advance. This includes determining what you're going to need to collect versus what might be available through partners or other sources and figuring out an approach for sharing those data. It also means developing plans, if needed, to create, pilot, and refine data collection instruments. This includes ensuring that the items are valid and the language in the data collection instruments is culturally and developmentally appropriate for the intended survey responders. And as I noted earlier, we strongly suggest that you consider the challenges of accessing hard to reach populations, for example, high mobility rates among those involved in foster care or justice involved youth. This has implications for the length of follow-up and the resources that you're going to need for follow-up.

Finally, we strongly recommend that you collect implementation data to support the data you collect for your impact analysis. And I'll get into this very soon, but more, but this is going to help you understand no only whether you observe impacts, but also why or why not? For example, if you find that there was ultimately minimal contrast between the treatment and comparison groups, you may be able to explain if you're unable to detect treatment affects, or, as another example, if you find subgroup differences in impacts, for example, by gender, you may be able to explain them by accessing information from focus groups or discussions with participants or program facilitators that might suggest that girls found the program more compelling than boys. So, as you can see through these prior examples, conducting an impact study without a companion implementation evaluation may get you to a place where you have study findings but you can't really interpret what they mean.

Now that I've hopefully sold you on the importance of conducting an implementation evaluation in conjunction with an impact analysis, you need to think about what data you might be able to collect to help you understand both the intervention and comparison groups experiences. This includes collecting data on fidelity, which is the degree to which the program was delivered as intended, with high quality dosage, or the degree to which the individuals attended the program and received the intended content and activities, feedback from key interested parties, this may include information from program participants, program leaders, facilitators, parents or caregivers, and community leaders, and can take the form of interviews, focus groups, or surveys. Intervention contrast, as we've previously discussed, including understanding the comparison groups' experience and the overall contrast which is critically important for interpreting program impacts. And we also suggest that you consider measuring implementation at the component level, particularly if you're considering conducting any analyses of components.

Finally, I'd like to touch only briefly on issues related to planning for analysis and reporting of findings. As you're planning a study, you should consider analytic methods that will present credible estimates of program impacts. Here I provide a link to the TPPER standards that I mentioned early, that provides guidance on producing credible estimates of impacts. We also suggest that you consider planning to apply Bayesian methods for interpreting impacts. This is particularly helpful for small subgroup analyses of hard-to-reach populations. OPA has resources available to think through these analytic concerns.

And regarding dissemination, we'd like to suggest that you think about dissemination from the very beginning, especially as it relates to keeping partners and key interested parties engaged and up to date on progress and challenges. Dissemination can happen throughout the lifespan of rigorous effectiveness studies. This could be done to support understanding among partners and community members that are part of the study, to support program improvement, and to support broader learning in the field. And we encourage you to think creatively about dissemination and how to use different methods of dissemination to tailor to different audiences.

So, I'd like to close this presentation by mentioning a number of resources available on OPA's website and on the RHNTC training site, that can support you in planning for an impact evaluation as well as any other formative or summative evaluation work that you might be considering. In particular, I want to call your attention to the impact evaluation toolkit, which includes a compilation of numerous resources that go into much more detail than I was able to cover today. The toolkit allows you to filter resources by different categories, including planning, implementation and data collection, analysis, and reporting, and we definitely encourage you to check that out. We also provide a link to a brief on calculating minimum detectable affects, impacts, and teen pregnancy prevention evaluations, which can be helpful as you consider the sample size that you might need to detect statistically significant differences in your primary outcomes.

And with that, I'd like to just say thank you, and we welcome any questions that you might have. And I believe we've already provided a link to these slides so you can have a link to all of these resources if you access the slides. And I'm going to turn it back to Tara to moderate the Q and A session.

[Tara Rice] Thank you Lauren, for that great presentation. So at this time we'll be taking questions. There's two different ways we can take questions, if you have a questions you can either enter it into the chat box and we'll read it off there, or if you would like to, you can raise your hand using the hand icon at the bottom of the screen and when I call on you, Rick can unmute you and you can ask your question directly. So at this time, any questions?

[Rick Stoddard] Douglas Taylor, you're unmuted.

[Douglas Taylor] Thank you. I believe it was on slide 12, you had a list of things to consider if you think you're ready to do an impact evaluation. And I kind of feel like we have an innovation that meets all of those expect the second to the last. And there was also a slide that, later, that talks about pilot studies. So I guess I'm just curious as to what do you really need to have in place in terms of evaluation of a program that says okay, we've collected enough data that says, well the type of data you need to collect, and that it's enough data to say yes, we're ready to move on?

[Lauren Scher] Yeah, I don't know if there's like a true answer to that question, because I'm sure that there are lots of, there's a lot of gray in this. But in terms of like you really do need to have some evidence that this is promising, right? And so maybe that means you're getting really, really close to doing a well powered study, but you still might need to do a bit more piloting and testing, you might want to be looking at like pre-post outcomes for the group that you're serving, you might want to do like a small scale type of pilot even to test, testing out your procedures for finding an appropriate comparison group. But I do think that you would not want to jump into this unless you're really getting a sense that the outcomes that you're expecting to see some changes in, that you're seeing that. Does that make sense?

[Douglas Taylor] Yes, thank you.

[Lauren Scher] That's why it's so important to do a lot of that formative evaluation back with over here in this, really thinking about doing outcome studies to see whether a program is assisted with favorable outcomes. And if you can, if you feel pretty confident about that, then moving onto that impact study makes more sense.

[Tara Rice] Thank you, Lauren. There's a question in the chat from Jonathon and I just want to say that we're not taking questions about whether, about specific proposals for any of the NOFOs, so those questions are not questions that we're going to be answering on this call, but if you have questions that relate specifically to the content of today's presentation, we will be happy to answer those questions. Rick, are we seeing any other hands raised?

[Rick Stoddard] No other hands at this time.

[Tara Rice] Okay. Well, if there aren't any other questions at this time, then I think we can close this webinar. This webinar is being recorded, so we will be posting the presentation, probably in about a week, on our RHNTC, the Reproductive Health National Training Center site. So, it will be available if people want to review it again. Lauren, are there any other closing thoughts you have?

[Lauren Scher] No, I don't have any other closing thoughts other than I just moved back to this resource page because I do think there's a lot of really great resources, I very quickly described a lot of different things that really get in-depth treatment in a lot of these resources that are here. So, I strongly encourage if you haven't looked at these resources, to dig into them.

[Tara Rice] And remember that you can download the slides if you haven't already done so, at the link in the chat so that you can get to those resources. So, thank you Lauren, and thank you everybody. If there are no other questions, then this concludes our presentation. And I hope that everybody has a great rest of their day.

[Lauren Scher] Thanks. Bye, everyone.