Local Evaluation: Answering Key Questions Transcript

Please stand by for real time captions.

Welcome

Hello everyone and thank you for attending today's webinar, Local Evaluation: Answering Key Questions.

Before we begin we wanted to cover a few housekeeping items. At the bottom of your audience console are multiple application widgets that you can use. You can expand each widget by clicking on the maximize icon on the top right of the widget or by dragging the bottom right corner of the widget panel.

If you have any questions for presenters during the webcast, you can click on the Q&A widget at the bottom and submit your question there -- tap at the bottom of the screen. We will try to address as many questions as possible during this event, but if a fuller answer is required or we run out of time, your question will be answered later through an FAQ document that will be posted on the websites for Responsible Fatherhood and Healthy Marriage and on the FaMLE Cross-Site website.

A copy of today's slide deck and additional materials are available in the resource list widget which looks like a green folder on the bottom of the screen.

If you have any technical difficulty, please click on the Help widget. It has a question mark icon and covers common technical issues.

A recording of the webcast will be available soon after the webcast and can be accessed using the same audience link that was sent to you to access this event.

Now I would like to introduce Robin Dion, Senior Fellow at Mathematica Policy Research. Robin, you now have the floor.

Introduction

Great, thanks Brice.

Welcome everyone to the last in a series of webinars presented by the Fatherhood and Marriage Local Evaluation and Cross-site project, also known as the FaMLE cross-site.

Sponsored by the Office of Planning, Research and Evaluation at the Administration for Children and Families, this webinar, like the others in this series, is specifically intended to provide assistance to organizations interested in applying for a new federal healthy marriage or responsible fatherhood grant. This includes the Healthy Marriage and Relationship Education grants, the New Pathways for Fathers and Families grants, and the Responsible Fatherhood Opportunities for Reentry and Mobility grants.

One of the exciting things offered by this round of grants funding is the opportunity for grantees to conduct their own local evaluations. Because this is a new feature of this round of grants there may be a lot of questions about the local evaluations. So the purpose of this webinar is to discuss the

expectations regarding local evaluation and to provide some information that you may want to consider as you develop your plans.

There will be several parts to today's webinar. To provide some context I will first describe what the FaMLE cross-site project is and how it relates to the healthy marriage and responsible fatherhood grants. Next we'll discuss some of the overall purposes of the evaluation and the different types of local evaluation that are allowed under the new funding opportunity announcements. We'll go on to discuss some of the pitfalls to watch out for in planning and conducting local evaluations and we will be providing tips on working with a local evaluator and Institutional Review Board.

We will have some time at the end to address at least some of your questions so please remember to use the Q&A box on your screen to submit your questions. My hope is that after today's webinar you will come away with a better understanding of how the funding opportunity announcements distinguish between collection of performance measures, local evaluation, and federal evaluation. We also hope you will gain a better understanding of the range of research questions that can be addressed by your local evaluation and the key characteristics of various types of high-quality evaluations.

Let me first introduce today's speakers, as Brice mentioned, my name is Robin Dion and I am the project director of the FaMLE Cross-Site Project. My colleague joining me today is Sarah Avellar the FaMLE Cross-Site Deputy Project Director and at the end of the webinar Seth Chamberlain and Julie Leis from the Office of Research, Planning, and Evaluation at ACF will address your questions.

So what is the FaMLE Cross-Site Project?

For those of you who have attended the other FaMLE cross-site webinars please bear with me over the next couple of minutes. We will get right into the new content after that.

The FaMLE cross-site project focuses on providing support for the 2015 cohort of healthy marriage and responsible fatherhood grantees sponsored by the Office of Family Assistance and ACF. This support will specifically be for collecting performance measure data and assistance for grantees' local evaluations. Ultimately, the project will also conduct analyses of data that will be collected across all of the grantees, which is why the term cross-site is included in the project name.

ACF first engaged Mathematica to conduct the project beginning in fall of 2013. The first activity was to examine, recommend and develop a comprehensive set of research-based performance measures that are intended to help grantees manage their programs and also provide high quality data on program operations and outcomes of participants. We also began development of the nFORM system, which will provided for free to grantees with all of the performance measures and reports already programmed in.

We also developed a website where you can access many resources for developing strong grant proposals. Once the grants are awarded, the project will continue to provide training and ongoing support to grantees as they use the web-based system to collect and report on performance measure data.

In the second major area of activity, the FaMLE cross-site project will also provide evaluation technical assistance and support for grantees as they conduct their own locally-led evaluations. As you know,

these evaluations are separate from the effort to collect and report on performance measures; local evaluations may be descriptive studies or they may be impact evaluations.

And third, to develop the big picture of all of the grants' achievements, the FaMLE cross-site project will be aggregating the performance measure data across all the grantees and conducting cross-site analyses in 3 areas: program design, implementation, and outcomes.

To help you get a better sense of the agencies at ACF and how the FaMLE Cross-Site project and the OFA grants are linked, this diagram shows how the organizations are related.

The box at the top of the diagram shows how ACF includes two agencies– the Office of Family Assistance, or OFA, in the left-hand box, and the Office of Planning, Research and Evaluation, or OPRE, in the right-hand box.

You can see that on the left that OFA oversees the grant funding for the responsible fatherhood and healthy marriage grantees. On the right, you can see that the Office of Planning, Research, and Evaluation (OPRE) oversees Mathematica's work on the FaMLE Cross-Site project, which will be providing support to OFA's grantees in collecting and reporting on data and conducting local evaluations.

The double-headed arrow in the center of the diagram between OFA and OPRE signifies the close collaboration that these two agencies have developed to facilitate high quality data collection and other research activities among grantees. For example, OFA has been highly involved with OPRE and Mathematica over the past year and a half in the development and refinement of the performance measures, developing the data collection modalities, and designing report formats.

Let's talk a little bit about the reason and the purpose for evaluating programs. You may have heard that evaluation is hard and that is sometimes the case. But good evaluation is a critical component of building a field of practice that is high quality, effective and sustainable. In general evaluation is really nothing more than a systematic way to determine whether a program is achieving the goals it has set out to achieve. It's also a way to address questions that are specific to certain program components and that if answered, could result in better programming. For example you can systematically test whether one program strategy results in better workshop attendance compared to another strategy. In any case, good evaluation uses methods that are objective and lead to unbiased results. Sarah is going to be talking about that some more.

One of the key benefits of the evaluation is that it helps programs identify the specific factors that have led to their success and helps them identify weaknesses that may need strengthening. Identifying the factors that are associated with successes and challenges can lead to valuable lessons that cannot only help your program improve but others as well. It can inform the next generation of programs and help advance the field. Finally, evaluation allows us to build the evidence base regarding the effectiveness of responsible fatherhood and healthy marriage programs. And a strong evidence-base, even if only in development, can ensure continued investment by funders and policymakers.

The acting director of the White House Office of Management and Budget recently put it pretty succinctly when summarizing the overall purpose of research evidence. He said, "since taking office the President has emphasized the need to use evidence and rigorous evaluation in budget, management, and policy decisions to make government work effectively. This need has only grown in the current fiscal environment. Where evidence is strong, we should act on it. Where evidence is suggestive, we

should consider it. Where evidence is weak, we should build the knowledge to support better decisions in the future." The goal of evaluation is not to arrive at a thumbs up or thumbs down, but to improve our decisions, our policies, and our practices.

As most of you know, the funding opportunity announcements state that grant applicants are expected to include a plan to conduct local evaluations, yet they also discuss the possibility that some grantees may be selected for participation in a federally-led evaluation. In the next few minutes I will try to clarify the differences between local and federal evaluations.

Let's begin by reviewing the expectations in each of the funding opportunity announcements with respect to local evaluation. Stated simply, a local evaluation is a research effort designed by the grantee and an evaluator selected by the grantee. It should be designed to address one or more specific research questions that the grantee is interested in. The type of local evaluation allowed under the funding opportunity announcements is determined by the level of funding as shown on this slide. Grantees funded up to $700,000 are expected to conduct a descriptive study. Grandees who are funded between $700,000 and almost $1 million are expected to conduct either a descriptive or an impact evaluation. Grantees who are funded between $1 million and $2 million are expected to conduct impact evaluations.

So the funding opportunity announcement discuss 2 general types of evaluations, descriptive and impact. There is a critical difference between these two types of evaluations that should be clearly understood. Descriptive studies describe something that the grantee is interested in which can be useful. What's important to know is that descriptive studies cannot establish cause and effect. Impact evaluations, on the other hand, are intended to address questions related to cause and effect.

Let me provide a quick example. Let's say a grantee wants to know if the number of clients with jobs increases over a period of time. This grantee could collect data on the client's employment status before and after the program. If there is a change the grantee could describe the change, but cannot claim it was caused by the program because many different factors could have affected that change. For example, a new employer may have moved into town with a variety of job openings so participants may have obtained jobs even in the absence of the program.

If the grantee really wants to know if their program actually increases the rate of employment among their clients, then they should plan an impact evaluation. The funding opportunity announcements require all impact evaluations to have a comparison group because that's the only way to know for sure whether any changes can be definitively attributed to the program and not to something else. Sarah will talk in more detail about why this is the case, but the point I want to make is that all impact evaluations must include a comparison group.

One version of an impact evaluation is a randomized controlled trial or an RCT. This is where you have a control group which does not receive program services. The funding opportunity announcements award bonus points for conducting an RCT because with this ability to offer conclusive evidence of program effectiveness, RCTs that are conducted correctly are the gold standard of evaluations that seek to establish causality.

Two final points here with respect to the funding opportunity announcements. All impact evaluations must also include a component that documents the programs design and implementation so that it is clear what exactly is being evaluated. This is also known as a process or implementation study. Finally, grantees are expected to work with an evaluator that is an independent from the organization that is

operating the program being studied. Independence is an important concept to assure that results are unbiased and objective.

Now let's turn to a discussion of how local and federal evaluations differ. The funding opportunity announcements describe two federal evaluations, Building Bridges and Bonds or B3, which will be an evaluation of a small subset of responsible fatherhood programs and Strengthening Relationship Education and Marriage Services also known as STREAMS, which will be an evaluation of a small subset of healthy marriage programs. Not all grantees will participate in these federal evaluations, in fact only a very small number will be asked to do so. All grantees must however agree to participate as a condition of their grant if they are asked to do so.

In contrast, all grant applicants are expected to include a plan for a local evaluation. As this chart shows a local evaluation is an effort that is specific to each grantee, is designed to answer a grantees specific research questions, and is led by a local evaluator. In contrast the two federal evaluations that I mentioned are conducted by large research firms who are under direct contract to ACF for these evaluations. These firms have long experience conducting multi-site rigorous evaluations of responsible fatherhood and healthy marriage programs funded by ACF.

Local and federal evaluations also differ in 2 other aspects, their funding and process. Regarding funding, all grantees are expected to reserve a portion of their total grant funding for conducting a local evaluation. The amount to be reserved is determined by the type of evaluation as stated in the funding opportunity announcements. Now after award during the planning period, some grantees will be selected for one of the federal evaluations. These grantees who are selected for one of the federal evaluations may receive additional funds for programming and study participation. When these grantees are selected for one of the federally-led evaluations, ACF may decide to incorporate the local evaluation into the federally-led evaluation or they may waive the local evaluation requirement.

Now grantees were not selected for one of the federal evaluations will continue to work with their local evaluators and the FAMLE cross-site team at Mathematica to refine their local evaluation plans during the planning period. Once the local evaluation plans are approved by ACF, grantees will implement their evaluation plan throughout the grant period.

Let me conclude by emphasizing that regardless of whether a grantee is involved with a local or federal evaluation, all grantees are still required to collect and report on the performance measures that are programmed into the nFORM system. These performance measures and the web-based system for collecting them were the topic of the past 2 webinars so if you missed them, please visit the FaMLE cross-site website to hear a recording.

Let me just summarize to be clear, all grantees must collect performance measures as shown in the nFORM system. All grantees are expected to propose a local evaluation, either descriptive or impact. After grant awards are made, a few grantees will be selected for one of the federal evaluations. And technical assistance will be provided to all grantees, whether they are collecting performance measures, conducting local evaluations, or participating in federal evaluations.

With that I'm going to turn it over to Sarah. She will discuss how to choose a research question and an appropriate method for planning your local evaluation. She will also discuss some of the key features of various evaluation designs and other factors to consider in planning and conducting strong evaluations. Sarah?

Research Questions

Thanks Robin. As Robin said I will be providing some guidance and recommendations for designing a local evaluation. The first step of designing an evaluation is thinking about what you want to answer.

Research questions are really the foundation of evaluation. They determine your research design, your data collection, and your analysis. The FOA stipulates that the proposed research questions must relate to the grantee's specific programming approach and that the questions must expand the research base. For example, if you're implementing an innovative approach, you should consider evaluating it.

Also consider what information would be useful to your staff and other stakeholders, including your partners. You can also talk to your local evaluator—and we'll talk more about this person later in the webinar—the local evaluator can also help you to come up with some research questions. But you want to select questions that are of interest to your team. Because it's just human nature: if you don't care about the answers, it will be very hard to put the effort into the evaluation.

Once you have a sense of what you want to ask, you next need to consider the evaluation design.

Evaluation can take different forms, and those forms answer different questions.

We have four designs in this table. The first two are impact studies, which are designed to assess the impact or effect of the program. The first is a randomized controlled trial, or RCT. Depending on how it's designed, it can answer whether the program as a whole or a particular component affects clients' outcomes. A quasi-experimental design or QED also answers whether clients' outcomes are affected by the program. But a QED is less rigorous than an RCT so the answer is not quite as clear cut.

The next type of design, a pre/post design, is descriptive. A pre/post addresses whether participants' outcomes change over time but cannot answer whether the changes were the result of the program.

Last an implementation study can fall into the either the impact or descriptive category depending on how it's designed. But it always examines the factors associated with or leading to high quality implementation.

Let's first dive a little deeper into impact designs, including their key features and why they're important.

A tricky concept is the difference between change and effect. There are many factors that can cause change. For example, someone may be motivated to change their circumstances. Or there is often natural change over time. A clear example is with children who naturally develop and mature over time. There also can be broader contextual or environmental influences, such as improvements in the economy or access to other services.

If we want to know whether a program caused a change we have to rule out all alternative explanations for the change. To do this, we need to know what would have happened without the program. If the program caused a change, then without it, the change wouldn't have happened. The impossible ideal for testing this is to have the same people simultaneously participate and not participate in the program. If it's the same people you can rule out any influences of personal characteristics; if it's the same time, then you can rule out any other contextual changes. But obviously you can't do either. So the realistic alternative is to have a comparison group, which is a group of

people who do not receive the services being evaluated. The comparison group is the stand-in for the program group not getting program services.

This is a really important concept, so let's unpack it just a little bit more. Suppose you have one group who receives services. This is equivalent to a pre/post design where client characteristics are measured sometime before and after the program. In this hypothetical example, we're measuring percent employed. You can see a big change from before the program, when 20 percent were employed, to after the program, when 50 percent were employed. Without any other information, you would logically conclude that the program had a substantial impact on employment.

However, let's now add a comparison group who didn't receive services. The comparison group also started with 20 percent employment, but actually had 55 percent employment around the same time as the end of the program. So this suggests that the program itself didn't actually improve employment but something else was operating. Maybe a new large employer came into the area, for example, or there was just a dramatic improvement in the community's economy overall.

As I said, the comparison group is intended to represent what would have happened to the program group if they had not received services. Remember, ideally and (impossibly) it's the same people in the program and comparison group. But since you can't have an alternate universe with the same people in both groups, it's critical that the comparison group be as similar as possible to the program group at the beginning. This is known as baseline equivalence. If the groups are similar at the beginning, then any later differences can be attributed to the program.

Here is another hypothetical example. Here we have a program group represented by the darker green bar and the comparison group with the lighter green bar. Say we've measured the percentage with high relationship dissatisfaction, which is obviously an unfavorable outcome. In this example, at the end of the study, a greater percentage of those in the comparison group are highly dissatisfied with their relationship, suggesting that the program group is better off than the comparison group, and potentially that the program had a favorable effect.

But now let's look at the groups at the beginning of the study. There were not initially equivalent on relationship dissatisfaction. The percent dissatisfied actually increased among those in the program group, but decreased among those who didn't receive the program. With this we can't make sense of whether the program had a favorable effect, an unfavorable effect, or no effect at all.

But if we tweak this example so that the groups were the same on relationship dissatisfaction in the beginning, we get a clearer answer. Here you can clearly see that the rate of relationship dissatisfaction increased for both groups, but it increased less in the program group. So the program prevented some people from becoming dissatisfied in their relationship, which is a favorable effect on clients.

As I mentioned earlier, there are two common designs for impact studies, randomized controlled trials or RCTs and quasi-experimental designs or QEDs. I want to cover two basic features of these designs. First, both designs have a program group that receives services and a comparison group that does not, but they differ in how the program and comparison groups are formed. In an RCT, people are assigned by chance, like by the flip of a coin. In a quasi-experimental design, the groups can be formed in different ways, but it's not a random process. For example, some people may receive the program because they happen to the live in the service area, whereas the comparison group could be made of people who live somewhere that don't access to the program. Or the program group is formed by

people who apply when there are openings and the comparison group is made up of those who apply when the program is at capacity.

The way the groups are formed affects the likelihood of baseline equivalence, the second feature highlighted in this table. With an RCT, because assignment is random, the groups have the same characteristics, on average. This is true whether or not you measure them. So for example, if 50 percent of people who enter the evaluation are highly motivated to change, then, on average, 50 percent of your program group is highly motivated as is 50 percent of your comparison group. This equivalence holds for any and every characteristic, which is one of the key strengths of the design.

In contrast, for a quasi-experimental design, because the groups were formed by a non-random process, you can't assume that the program and comparison groups are equivalent. You have to measure characteristics at baseline and determine whether the groups are the same. Unlike groups formed randomly, you can't assume that the groups are the same on any characteristics that you haven't measured. So you can only determine baseline equivalence for the characteristics you measure and analyze.

Because RCTs are more rigorous than QEDs, we do suggest using them if possible. If you want to know whether your program is effective, random assignment is one of the best methods for answering the question. Other designs can give you the wrong answer. Because of its advantages, random assignment is used not only in a wide range of studies, but in high stakes evaluations, such as testing the effectiveness of medicines or medical treatments.

Another question to ask yourself is whether there are people in the community in need that you can't serve? If the answer is yes, than random assignment is a fair way to allocate services that can't be offered to everyone.

However, say you can't use random assignment but you want to do an impact evaluation. What are your alternatives for forming a comparison group?

First, you could draw a comparison group from a geographic area which doesn't have similar services to yours, such as a nearby county. You do want the county to be the same as your county in other ways, such as other available services and population characteristics. You will also have to consider how you will reach people in the county for them to consent to the evaluation and for you to collect data.

Second, you could partner with an agency that serves a similar population as yours but does not offer similar services. Clients of that agency who consent to the evaluation could be part of the comparison group. However, it is important that the other agency's services are different enough from your program, or it's going to be very difficult to detect any differences or effects from your services.

Third, if your program typically reaches capacity, you could form a comparison group from those who cannot be served because the program is full. If program capacity is the only reason that someone cannot be served, this is a very good impact design. But to form a comparison group, you will have to recruit just as much when the program is full as when there are openings.

And fourth, you could use administrative data. For example, if other agencies served similar populations and collected data, you could develop a data sharing agreement with them. You also could try to obtain state-wide or national data on characteristics such as child support or wages. But there

are often many steps to obtaining such data, so you would likely need to work with an independent evaluator with experience acquiring such data.

We also have some tips on strategies you should avoid when considering how to form a comparison group. First, don't create a comparison group from people who agreed to participate in your program but either never showed up or quickly dropped out. The key reason is that these people are likely to be different in some important ways than those who did show up. Maybe it's something like underlying motivation or barriers in their lives that interfered with their attendance. These types of differences may create bias, which is erroneously shifting results in one direction or another. With biased results, you may conclude your program had effects when it did not, or you may conclude your program did not have effects when it did.

Second, we caution you from forming a comparison group from those who are not eligible for your program or who were otherwise not a good fit. For example, if you routinely screen out those who say they are not interested in employment, do not use them for the comparison group to evaluate your program's effect on employment. Again, underlying differences may bias your results.

Now that we've discussed impact designs, let's turn to descriptive pre/post designs.

Typically in a pre/post design, clients' characteristics are measured before they participate in a program and after. This design does not have a comparison group, so it can't answer whether any changes you measure were caused by the program or other factors. It can, however, show how clients' changed over time. For example, you may examine whether couples' relationships improved over time or clients earnings increased. Pre/post designs are a good first step if more rigorous designs are just not doable.

And then we have implementation studies. These actually may be designed as descriptive studies or to evaluate impacts.

Implementation studies examine factors that lead to program outputs, such as clients' participation in the program, staff turnover, or fidelity to the curriculum. Often an implementation study can identify ways that a program can improve program operations. An implementation study can be designed as either a descriptive or impact study. Let me give you two examples of an implementation study that focuses on client participation. The first implementation study is an impact study. In it, clients are randomly assigned to either participate in a program that is offered in 6 one-hour sessions or to participate in a workshop that offers the same content in 2 3-hour sessions. This study could determine whether participation is better in one format versus the other.

In the second implementation study, staff conduct focus groups with client to learn about why they did or did not attend. Although this study does not support any causal conclusions, it may suggest ways to modify the program that could improve participation.

That is an overview of study design, now I'm going to turn to other evaluation factors that should be considering when designing an evaluation.

As I've just described an evaluation design determines what can be learned, but how that evaluation is executed determines what will be learned. Evaluations rarely go exactly as planned, so we suggest you anticipate common problems and think beforehand how you might address them. I'll briefly go through each of these seven factors listed on the slide.

Many programs have multiple services, and by design, some clients may receive certain services, but not others. So it's important to define the specific services that will be evaluated. For example, are there core services that all or most participants receive? If so, those services may be the best option for evaluation because you can include most clients in the evaluation and they're receiving a relatively consistent set of services. But you also may want to include supplemental services or services from partners in the evaluation, which you can do, but it just requires a little more thought. For example what percentage of your clients typically receive those services and then can be included in the evaluation?

In an impact study, you are assessing the effectiveness of services offered to the program group that are not offered to the comparison group. A rule of thumb is that the greater the contrast in services, the more likely it is to see impacts. So if the comparison group receives services similar to those the program group receives—either through your program or other programs in the community—the evaluation can't measure the full impact of all services.

Before any client can participate in an evaluation, they must be fully informed about what they will be asked to do as part of the study, the study's risk, and where to go with questions. Then, with this information, they must consent before they can be included in the evaluation. An institutional review board or IRB will need to approve the consent process. Typically the process includes a consent form, which conveys the information in writing, and having a trained staff person who goes over the form with the potential client. Keep in mind that not everyone will consent to the evaluation, so you should plan ways to increase the likelihood of consent. This can include being prepared to address common concerns that clients may have. It can also be helpful to convey to the client the importance of the study to the field.

Attrition is a factor that can substantially increase the risk of bias in an evaluation. Attrition is basically missing data. Sometimes attrition is used to refer to clients leaving the program or dropping out of services. But here I am referring to attrition from the evaluation. And as I will discuss in a moment, a client can leave the program but still be part of the evaluation. Attrition from an evaluation means someone in the program or comparison group does not provide data, for example, not responding to a follow-up survey.

In this diagram, the top panel represents people who have been randomly assigned. The blue and red characteristics are equally distributed across the groups. As you'll recall, this is a benefit of random assignment: the groups are the same on average, on all characteristics. However attrition can erode this benefit.

For example, say the blue and red represent motivation to change and that blue people are less motivated and red people are more motivated. Although the groups are the same at the beginning, through attrition, the program group has ended up with an equal number of motivated and unmotivated people, but the comparison group has ended up with a majority of highly motivated people. They may actively seek out other services or make other changes on their own. At the end, when outcomes are measured, it may look as though your program is less effective than it is, because the groups were dissimilar on this underlying characteristic.

Obviously attrition is a problem, so it should be minimized as much as possible. For an evaluation, data should be collected on everyone. So for example, say some of the clients assigned to the program

group dropped out of services or never attended. It's important to still collect data from them, because otherwise they count as attrition.

I'll just point out that the performance measures, which will be collected during program services, are not measuring effects of the program because they are only collected from those who are attending services.

It can be difficult to track people over time, especially if they disengage from the program, and it's often more difficult to collect data from members of the comparison group who aren't involved in program services. Therefore it's important to collect as much detailed contact information as possible at intake. You will want to collect multiple phone numbers, for example, and contact information about other people who might know the client's whereabouts if you can't get directly in touch with him or her.

As I mentioned earlier, baseline equivalence is critically important for an impact study. We can't assume that most designs have baseline equivalence. Any quasi-experimental design in which the program and comparison groups were not formed randomly can't be assumed to have baseline equivalence. And any randomized controlled trial with attrition—which is almost all of them—cannot be assumed to have equivalence if data are missing from some sample members.

If you can't assume baseline equivalence, that means you have to assess whether the program and comparison groups were similar initially. Of course, you can only know the ways the groups are similar based on what you measure at the beginning of the study. You'll need to consider what you want to measure at baseline so you can check for equivalence on those characteristics. Common baseline measures include why someone is one group or the other. And you might be wondering what this means. So let's again consider motivation. Suppose you are doing a quasi-experimental design. The people who seek out program services form your comparison group. They might be very motivated to improve their parenting, their relationship, or their economic situation. That's exactly why they are pursuing services. If you draw your comparison group from another set of people who don't seek out services, they may not be at the stage yet where they are ready to or interested in change. With a highly motivated program group and a less motivated comparison group, you are likely to bias your impact study. That's because motivation is probably related not only to what group they ended up in but their outcomes as well. Ideally, you select the comparison group so you think they would be similar on such characteristics as underlying motivation. And in this example you would also want to measure motivation to change at baseline so you can assess the group's equivalence on this factor.

Other important baseline measures include demographic characteristics, such as age, race and ethnicity, and education; and baseline measures of the outcomes of interest. That is, you usually want to measure at pretest anything you are measuring at posttest.

Generally speaking, the larger the evaluation's sample size, that includes both the program and comparison groups, the smaller the effects that can be detected. This is also sometimes referred to as statistical power. If you have low power, you probably can't detect small or moderate effects. Or to put it another way, if you have a small sample, the only effects that you will probably be able to see statistically have to be large. And large effects are very hard to achieve for any program because change is hard.

It follows that you want to maximize your sample size, that is, include as many people as possible in your evaluation. Keep in mind, maximizing the sample size involves getting people in the sample and

then collecting data on all of them. Sample size is affected by a number of factors including, how many people are interested in and eligible for your program, the number of people the program can serve, the number who consent to the evaluation, and the number for whom data are collected. When estimating how big your sample will be, you need to think about all the points someone can come in or drop out of the evaluation.

Another rule of thumb is that impacts will likely be larger with high participation. We assume that the more services someone receives, the better. However, low participation is a very common challenge. So when thinking through the evaluation, it's also important to be realistic about average participation rates. Consider ways to increase the likelihood that someone in the program group takes part in any service and receives a substantial amount of services, such as at least half of the workshop sessions.

The measures you choose also determine what you learn about your program. Selecting good measures is an art and science. For example, some outcomes that we're interested in have socially desirable answers or seem like there is a right or better answer. That is, people are probably not going to answer very honestly if you ask something like, "do you mistreat your partner?" This is an obvious example but there can be more subtle ways for a question to be leading, so that's something you'll want to keep in mind when selecting or creating measures. You also want to avoid jargon or confusing terms.

If you're using an existing measure, you'll want to consider features such as whether it's free and whether it's been used with populations similar to those you will be evaluating. Generally you want to select measures that have been tested with a population similar to yours because there may be cultural differences in how language or concepts are interpreted for example. Existing measures sometimes have indications of quality, including what's known as reliability and validity. These are empirical measures of how consistent a measure is and how well it captures the underlying construct.

I'll also note that you can use performance measures as some of your evaluation's measures. Those performance measures have been selected taking into account the features I just described.

In addition to the grantee staff, I'll just briefly describe two other parties who are likely to be involved in a local evaluation.

The FOA requires that the local evaluation be conducted by an independent evaluator. This means that the evaluator does not have a conflict of interest with the program or agency. An independent evaluator may come from an outside research organization or a local university, for example.

When deciding who to work with, consider whether they have done evaluations similar to what you hope to do. For example, if you want to do a randomized controlled trial, have they conducted those before? Ideally, the evaluator not only has experience with similar research designs but also has working with similar programs, so he or she has substantive expertise as well.

Once you've selected an independent evaluator, figure out how you'll divide up responsibilities. For example, the evaluator may conduct random assignment, but grantee staff tell the clients the results.

An institutional review board, or IRB, is an independent committee that reviews and approves the research. The IRB is responsible for assessing the risks to participants in the evaluation. I mentioned earlier, for example, that the IRB must assess the consent procedure to determine whether they think the process will fully inform someone before they decide whether to be in the evaluation. Universities

typically have IRBs and there are also freestanding IRBs that work for a fee. The FOA indicates that most grantees will either need IRB approval or a waiver. Applicants must complete the Protection of Human Subjects Assurance Identification/Certification/Declaration of Exemption form. The instructions are in the FOA, but briefly, most applicants will check the 3rd option in box 6, and then take further steps if awarded a grant. Again, the instructions are in the FOA, but this slide shows the website address and a screen shot of the form I'm referring to.

As I mentioned, most applicants will select the 3rd option in box 6, shown here. It might be a little bit hard to see but these slides will be available. It states that the applicant will provide an Assurance and Certification of IRB review and approval upon request. I know that I have just presented a lot of information. So I want to reassure you that we have other resources for you to use. With that I am going to turn it back to Robin.

Robin?

Resources for Grantees

Thank you Sarah. Sarah has just provided us with a crash course in program evaluation and raised some of the common challenges that arise. She's also described many of the underlying principles of good evaluation and provided some general recommendations and best practices. Still there isn't always a right answer for every situation. Each grantee and local evaluator must work closely together to decide what will work best for their programs following the principles of good evaluation design. To further help, Mathematica will provide grantees with technical assistance that is specifically focused on their local evaluations under the FaMLE Cross-Site Project.

As you develop your grant application and plans for local evaluation, you are encouraged to visit the FaMLE cross-site website which includes a great deal of useful information regarding this subject under the evaluation design tab. It includes questions for you to consider as you develop your plan. It provides targeted tips and offers links to other resources that might be useful. It also includes a flowchart that illustrates what type of design you should consider based on the type of question you would like to address.

Before we wrap up let me remind you that this is the final webinar in our series. We conducted three other webinars on program design, program measures, and the nFORM system. All the webinars are recorded and will soon be available on the FaMLE cross-site website. I would like to draw your attention to one more webinar that will be conducted by the Office of Family Assistance to describe the two federal evaluations--the B3 and the STREAMS evaluation. You may be interested in attending that webinar as well.

Hopefully all of these resources will help you begin to address your questions. In addition we will take some questions today and we will be recording and posting the answers to the questions that you raised both during today's webinar and during the other webinars on the FaMLE cross-site website. This includes any questions that were submitted during the webinars but which we didn't have enough time to verbally answer. As before, please understand we can't answer questions outside of this series of webinars because everyone needs to have access to the same information to ensure a fair grant competition. Thank you so much for your attention today. I will now turn the webinar to staff from OPRE who will address your questions.

Seth?

Question & Answer

Thanks Robin. My name is Seth Chamberlain and I'm from the Office of Planning, Research, and Evaluation. Thank you all for your participation today. I will be reviewing questions that have been submitted as part of this webinar but I will not be able to address every question, just as I haven't been able to address all of the questions from the other webinars. We are reviewing every question and developing FAQs that will be posted to the FaMLE cross-site website. We believe by next week. That way everyone can have access to the same information. So thank you for your questions--let's begin.

The first question is: what if the independent evaluator is a faculty member but the University is not local? That's a good question.

If the local evaluator is not geographically local, that is okay. The term local evaluator is simply used to denote someone who is not a federal evaluator. The local evaluator could be somebody from a university that is in town, in state, or a university out-of-state, or research organization that is in town, in state, or out-of-state.

We also received some other questions related to what if a grant is being -- I apologize, I don't have the question right here in front of me, but I saw it coming in and I want to address this issue. There is a question: what if a grant applicant is a department in a university, can another department in the University be the local evaluator?

In general yes, but only if the grant applicant, and if awarded, the grantee, demonstrates that there is true independence between the program and the evaluator, and if safeguards are put in place to ensure there is continuing independence of the evaluation. For example, the implementing and evaluating groups must hail from completely different schools within the University.

There is another similar question that has come up which is: can a local evaluator be someone that is in-house to the organization that is implementing the program?

The answer is generally no. The organization may have a vested interest in the results of the study and even the results of a descriptive study. Unless the grant applicant, and if awarded, the grantee, demonstrates true independence between the program and the evaluator and if safeguards are put in place to ensure continuing independence of the evaluation. For example, the evaluating group must demonstrate there will be no financial benefit to it or to the organization from the results of the study.

The next question is: if our impact evaluation begins in year two and if we have adequate sample after two years, do we need to continue recruiting beyond year three or could years four and five be used for reporting of findings and follow-up?

The answer is that a lot of these decisions will be made post-award. The sample sizes that will be needed should be proposed in the applications and justified. But the exact rollout of the evaluation, the exact timing of the analysis of data, and the publication of findings will be local evaluation determined, or determined by the specific local evaluation.

The next question is: do we need to use nFORM and the required tablets for impact evaluation?

To refresh everyone's memory, or if you did not attend yesterday's webinar on the management information system that has been created for this cohort of responsible fatherhood and healthy

marriage grants, the nFORM system is the management information system that has been created for this cohort of grants. There are surveys for applicants which they must complete using the nFORM system. And then there are surveys, a pretest and posttest, for participants to be taken using the nFORM system. So the question is: do we need to use nFORM and the required tablets for impact evaluation? The tablets being the computers that participants or applicants use to complete the nFORM surveys. The answer is: it depends. It depends what the design of the local evaluation is. To be more specific about this I'd like to draw everyone's attention to page 31 of the healthy marriage funding opportunity announcement and page 33 of the responsible fatherhood funding opportunity announcement. Under the section that's entitled funded activities evaluation plan, the second paragraph says more specifically applicants must propose a descriptive local evaluation plan or an impact local evaluation plan in accordance with the funding levels requested, which will answer one or more grantee-specific research questions. So all applicants must propose one descriptive local evaluation plan or impact local evaluation plan which will answer one or more grantee-specific research questions. Depending on what those grantee research questions are, the research design and data collection associated with the design should be proposed. One could imagine that performance measures in nFORM would be part of a local evaluation. One could also imagine different local evaluations without those measures. I want to draw everyone's attention to another section in the same place in the FOAs. On page 31 in the healthy marriage FOA and page 33 -- I'm sorry page 34 in the responsible father FOA. Under the bullet called research design, the second bullet under research design says applicants must include a justification for why the proposed research design is best suited to answer the research questions. So that means nFORM and the tablets could be proposed to be used for the local evaluations and they will be set up so that control or comparison group members can complete the applicant characteristics and the pre-and post-tests. However, applicants must include justification for why the proposed research design is best suited to answer the research questions.

The next question is: when you talk about the need for fully informed consent, you said this was true for participants in impact evaluations. Wouldn't this be true if collecting outcome data for a descriptive evaluation too?

The answer is: yes. Informed consent will be needed for all of these folks. In short, informed consent is needed any time human subjects are involved in research.

The next question is: is there a minimum or recommended sample size for the impact study?

The answer is: no. Applicants, as I just mentioned, should include a justification for why the proposed research design is best suited to answer the research questions. I would like to draw everyone's attention in the same section of the FOA to the bullet entitled research design but this time I will talk about the first paragraph in that bullet. Under research design, it says applicants must propose a specific research design in their plans including details regarding staffing, timeline, recruitment of participants, planned sample size etc., etc., etc. That means that the sample size will depend on the research questions in the research design that is proposed.

The next question is: In random assignment could the treatment group receive the full healthy marriage 16-hour curriculum plus a five-hour job readiness class? And there's another question in this question. Could the comparison group receive the job readiness program if it is their only service and healthy marriage was not provided?

The answer to this question: I want to draw everyone's attention to another section of the FOAs. The FOAs have an excellent example of some of the range of research topics and designs possible. I'm looking in the healthy marriage FOA on page 10 under local evaluations and I'm looking in the responsible fatherhood FOA under local evaluations. I should mention when I'm talking about the responsible fatherhood FOA and the healthy marriage FOA, the language is nearly identical in the ReFORM FOA. Under the section local evaluations, the third paragraph starts: the proposed research questions must relate to their specific programming approach and that will expand the evidence base. Then there are 4 bullets of example priority topic areas. I want to highlight these. Topic areas for the local evaluations include recruitment and program participation, programming, program supports, and overall program outcomes. To answer the specific question, there is an example that is listed after this bulleted list. It says for the topic areas listed here, descriptive studies would not have a control or comparison group. Impact studies would have a control or comparison group. For example a program offering three types of services could look for associations between participants' use of those services and outcomes. This would be a descriptive study. On the other hand, the program could randomly assign participants to be eligible to use one, two, or all three services and then analyze whether the groups had different outcomes. This would be an impact study. So while I'm not going to address the specific question that was asked because we're not going to assist any specific applicant with their application, what I will say is that the FOAs have an example of a local evaluation design where a program could randomly assign participants to be eligible to use one, two, or all three services and then analyze whether the groups had different outcomes and that would be an impact study.

The next question is: if we're doing an implementation study that is an RCT, randomized control trial impact study, should we conduct a second RCT or randomized controlled trial on a different question?

I'm not completely understanding the question. The question starts if we're doing an implementation study that is a randomized controlled trial, so a randomized controlled trial could have component that is an implementation study. But an implementation study doesn't by definition have a randomized controlled trial associated with it. But I would like to do is I would like to answer this question in two ways. The first is the question asks should we conduct a second randomized controlled trial on a different question and the answer is no. The funding opportunity announcement asks applicants to propose one descriptive local evaluation or impact local evaluation. I do want to draw everyone's attention to what is required for the proposed impact local evaluation. To answer this, I am going to draw everyone's attention back to the funding activities section of the funding opportunity announcement. I'm looking in the healthy marriage funding opportunity announcement at page 31 and I'm looking in the responsible fatherhood funding opportunity announcement on page 33. The third paragraph under funded activities evaluation plan says applicants that propose to conduct an impact evaluation must have and describe a comparison group who does not receive the services of interest and that is comparable at baseline, i.e. before a program begins, to those who participate in the service program. In addition to evaluation of impacts on specific outcomes, applicants proposing impact evaluations must also include an examination in close detail of the function and form of different parts of the service process. That means an implementation study or a process study. Just to reiterate, for applicants that propose to conduct an impact evaluation, they must also propose to include an examination in close detail of the function and form of different parts of service process. To go back to the example that we talked about before where a program could randomly assign participants to be eligible to use one, two, or all three services, note that when you are randomly assigning participants to be eligible to use one, two, or all three services that means that some participants would not receive all services of interest. It doesn't mean that all participants -- it doesn't mean that there are some participants in the program who receive no services. It's possible. That is a

possible design but it's not a required design in a randomized controlled trial. Again, a program could randomly assign participants to be eligible to use one, two, or all three services and then analyze whether the groups had different outcomes. This would be an impact study. I also want to say that in this example here, if we're doing an evaluation where the program randomly assigned participants to be eligible to use one, two, or all three services, depending on the services that are being analyzed in the impact study, those are the services that would need to be examined in close detail. That is the implementation or process study that would need to complement the impact study.

I'm going to check with my team and ask them over the team chat to confirm that my voice is coming in and that I'm not garbled or anything like that.

The next question: since the lowest level of funding only requires a descriptive local evaluation, will bonus points still be awarded for an applicant in that range that conducts an impact evaluation with a randomized controlled trial?

To answer this question, I want to first speak to the healthy marriage and responsible fatherhood funding opportunity announcements and then I want to talk separately about the ReFORM funding opportunity announcement. To discuss the healthy marriage and responsible fatherhood funding announcements I want to draw everyone's attention back to the earlier parts of the FOA. I'm looking in the healthy marriage FOA on page 10 and in the responsible fatherhood FOA I'm looking at page 12. Under local evaluations there are three bullets, the first set of bullets. The first says grantees are expected to conduct either a descriptive or impact local evaluation as follows. For the first bullet, grantees funded from $350,000 to $699,999 per year are expected to conduct a descriptive local evaluation. The second bullet says, grantees funded from $700,000 to $999,999 per year are expected to conduct either a descriptive or impact local evaluation. The last bullet says grantees funded from $1 million to $2 million per year are expected to conduct an impact local evaluation. So for the lower tier, those grantees are expected to conduct a descriptive local evaluation. For the middle tier, they are expected to conduct either a descriptive or an impact local evaluation, and for the top-tier they are expected to conduct an impact local evaluation. Those are the permitted local evaluation designs for those tiers of funding. Those are the permitted designs. That means other designs are not permitted. I will also say that the funding opportunity announcements have a possibility of five bonus points for applicants that propose randomized controlled trials that meet specific standards. Let's talk about that for a second and again, I will talk about ReFORM separately in a second. If you are considering applying for ReFORM -- please hold on for one second while I finish this discussion.

For the bonus point I'd like to draw everyone's attention in the healthy marriage funding opportunity announcement to page 49 and to the responsible fatherhood funding opportunity announcement to page 52. Five bonus points are available for those that propose impact evaluations related to randomized controlled trials. And it says that the following criteria, and this is what people who are reviewing grant applications will consider in possibly awarding the five bonus points, it says, the following criteria will be applied to applicants that propose local impact evaluations with a randomized controlled trial. Only those applicants meeting all criteria will receive bonus points. There will be no reward of partial points. The three criteria that must be met first, it must be a randomized controlled trial with relevant research question or questions. Second, there must be random assignment. It's a randomized control trial so there must be random assignment. The application must discuss the method and timing of random assignment and the comparability of research groups. Lastly, the applications must discuss the sample. That means the unit of analysis and target populations and the sample size and the methods to promote sufficient program participation. I should mention in the

healthy marriage FOA that methods to promote sufficient program participation should be a letter C under the B sample size. In the responsible fatherhood FOA it is a letter C under the letter B. But it is in the FOA so it will be used as part of the criteria for those bonus points.

Now I would like to talk about the ReFORM FOA. The expectations for local evaluations for ReFORM are different. ReFORM, for those that aren't immersed in this language, ReFORM is the Responsible Fatherhood Opportunities for Reentry and Mobility Grants Program. On page 11 of the ReFORM funding opportunity announcement there is a section that is called local evaluations. The second paragraph says grantees are expected to conduct either a descriptive or impact local evaluation. They should allocate at least 10% and no more than 15% of the total annual funding for the local evaluation. Please note that there are no tiers of types of local evaluation for the ReFORM FOA. Therefore, regardless of the amount of funding proposed in the application, applicants for ReFORM can propose descriptive or impact local evaluations. The criteria for judging those local evaluations is very similar to the criteria to judge the local evaluations that are proposed in healthy marriage and responsible fatherhood. In fact they are identical except there are no bonus points available for the ReFORM applicants.

The next question is: does the amount that has to be allocated for a local evaluation for fatherhood grants include the cost of laptops, program staff time to collect and enter data, share data with the evaluator, etc.?

What I'd like to do to answer this question is I'd like to direct everyone's attention to another section of the funding opportunity announcements. I'm in the healthy marriage funding opportunity announcement on page 35 and I'm in the responsible fatherhood funding opportunity announcement on page 38. This is the section called budgeting for evaluation. The title is not on the responsible fatherhood announcement but it's in on the page before on page 37. Under budgeting for evaluation it says the applicants' overall line-item budget and budget justification must also include detailed allocations for the range of required performance measure data and evaluation activities. Let me say that again. Detailed allocations for the range of required performance measure data and evaluation activities. That includes the local evaluation activities. And that includes the following: the first bullet says collection of performance data including cost of staff training, time to collect data, second bullet says storage of performance data including desktop, laptop, tablet, and purchase for ACASI online characteristics and pre-and post-test including headphones and maintenance, cost for staff to conduct regular activities like data entry, quality checks for reliability training, coding, etc. and monitoring performance data including cost for staff to analyze data, create and review reports, and plan to monitor adjustments. A lot of this language is specific to performance data however the first line of this part of the funding opportunity announcement says the applicants' overall line-item budget and budget justification must also include detailed allocations for the range of required performance measured data and evaluation activities.

Next question is: does the comparison group need to be selected from the same community as the service area?

The answer is: perhaps but not necessarily. It depends on the research questions being asked and the research design being proposed. As I said earlier, applicants must include a justification for why the proposed research design is best suited to answer the research questions.

Next question is: will quasi-experimental evaluations meet the threshold for impact evaluations for top-tier grants, even though it is not as rigorous as random assignment?

I'm not sure what a top-tier grant is. I think top-tier grant means those funded at the top-tier, however regardless of whether the grantee is funded at the second tier or the top-tier, those are the 2 tiers that may propose impact evaluations, regardless of whether they are funded at the second or the top-tier the local evaluation will be judged on its own merit regardless of whether it is an impact randomized controlled trial or a quasi-experimental design. It will be judged on its own merits.

The next question is: for the healthy marriage relationship education grant, if we choose to do six out of eight allowable activities to provide a broad array of services can we focus the impact evaluation on one or two allowable activities or do we have to do the impact evaluation on all six allowable activities?

To answer this question I want to bring everybody back to the first part of the funding opportunity announcements that talked about the local evaluation. I'm looking on page 10 in the healthy marriage funding opportunity announcement and on page 12 of the responsible fatherhood funding opportunity announcement. Under local evaluations there is a bulleted list of priority topic areas. The second bullet says whether certain program components or program structures, variations in intensity and duration of programming, or modifications to increase cultural competency are linked to better outcomes for participants. The paragraph after the bulleted list talks about an example where a program can randomly assign participants to be eligible to use one, two or all three services. Note that there is nothing in the funding opportunity announcements that says that an applicant must propose to evaluate the entire program, although evaluating the entire program is permissible. There is nothing in the funding opportunity announcements that says the entire program must be evaluated with a local evaluation. It must be related to the specific programming approach and that is one of the sentences in the FOA in this section, it says the proposed research questions must relate to their specific programming approach and that will expand the evidence base. But there's nothing in the funding opportunity announcements that says the entire program must be considered in the local evaluation.

We're out of time but as I mentioned earlier, we will be developing a frequently asked questions document which will include all of the elements of the questions that have been submitted and answers to them. And we hope to post this FAQs document next week. I anticipate it will be posted next Friday so I don't think you should look for it on Monday or Tuesday of next week. I do appreciate all of the questions. These are great questions. I look forward to answering them over the next few days in narrative form.

I do want to mention, as Robin did, that there will be a webinar tomorrow on the federally-led evaluations that my office, the Office of Planning, Research, and Evaluation, will also be putting on. I look forward to seeing you all again tomorrow. Tomorrow's webinar, today's webinar, and all of the other webinars associated with performance measures, nFORM, local evaluations, and the federally-led evaluations will be posted to the FaMLE cross-site website.

Thanks everyone and now I'll turn it back to Brice.

Thanks everyone for joining today. This concludes the webcast for today. Thank you and have a nice day.