David Murray:
Hello, my name is David Murray. I'm the NIH Associate Director for Prevention and Director of the Office of Disease Prevention. I want to welcome you to part seven of our course on Pragmatic and Group Randomized Trials in Public Health and Medicine. Part seven will cover alternative designs to evaluate multi-level interventions. This is part of a seven-part self-paced online course that's free and presented by NIH. We provide the slides for each of the modules, readings for the entire course, and guided activities for each of the components. Our target audience includes faculty, post-doctorate fellows, and graduate students who are interested in learning more about the design and analysis of group randomized trials.

We also want to reach program directors, program officers, and scientific review officers at the NIH who are interested in learning more about these designs. Participants should be familiar with the design and analysis of individually randomized trails and with the concepts of internal and statistical validity, their threats and defenses. We'd also like participants to be familiar with linear regression, analysis of variants, and co-variants, and logistic regression.

Our learning objectives are shown here. We expect that participants will be able to talk about the distinguishing features of group randomized trials and individually randomized group treatment trials and contrast those to individually randomized trials. We expect that you'll be able to talk about the appropriate uses of these signs in public health and medicine and for group randomized and individually randomized group treatment trials to discuss the major threats to internal validity, to statistical validity, the strengths and weaknesses of design alternatives and analytic alternatives, and to perform power calculations, or sample sized calculations, at least for simple group randomized trials. Finally, we expect you to be able to talk about the advantages and disadvantages of alternatives to group randomized trials for the evaluation of multi-level interventions, and that in fact, is the focus of today's presentation.

The organization of the course is shown here: we are on part seven, alternative designs. So, what about alternative designs? Do we have to use group randomized trials or individually randomized group treatment trials? People often complain to me that I've made their lives more difficult because group randomized trials and individually randomized group treatment trials are difficult and complicated and big and expensive, and do they have to use those methods?
Well, a number of methods have been proposed as alternatives, and it's only fair to talk about them, and talk about their strengths and weaknesses as we have for group randomized and individually randomized group treatments trials. We published a paper in 2010 that reviewed many of these designs that are listed here, and these are the ones that I'm going to talk about. I would also refer you Will Shadish and Tom Cook's book, "Experimental and Quasi-Experimental Designs for Generalized Causal Inference," published in 2002 that covers many of these designs.

Let's start by talking about multiple baseline designs. So, in a multiple baseline design, we usually have just a few groups. Let's imagine, for example, that we have four communities. And the intervention is introduced into these four communities one-by-one on a staggered schedule. Measurement is conducted in all of the groups, at each of the transition points, and, as I said, this kind of design is often used with just a few groups, three or four. Data are examined for changes associated with the introduction of the intervention is each group.

This is a figure that shows a hypothetical example of a multiple baseline design. All of the little circles are data collection points, and you can see that there are multiple data collection points before the introduction of the intervention in first community. There is an apparent intervention effect, just looking at the shift in the bubble of the line, and it seems to hold steady over the course of time. Sometime later, the intervention is introduced in community B. We see a similar shift that is consistent and holds over time. At some later point, the intervention is introduced in community C and later in community D. And in each case, in this hypothetical example, we see a fairly steady pattern and similar pattern before the intervention is introduced, then a shift and a different but steady and consistent pattern after the intervention is introduced. If you are so fortunate to have this pattern of results in your data, you can write a paper and claim an intervention effect, because we can all see if there on the page.

The evaluation of multiple baseline designs relies on logic, however, rather than statistical evidence. So as I just walked through the picture, that's the kind of conversation you'd be having in the discussion section of your article, writing up this study. You're looking for replication of the same pattern in each group, coupled with the absence of that kind of change otherwise, and if you get that, you interpret that as evidence of an intervention effect. If you just have a few groups, you have very little or no power for a valid analysis, such as a mixed model analysis or random co-efficient analysis, in this case. This kind of design, the multiple baseline design, is actually a very good choice if you expect to have large and rapid effects, but it's a very bad choice if you expect to have small or gradual effects. And it's a terrible choice if you think the intervention effect may vary from community to community or from group to group. The pattern that I showed you in the hypothetical example had a very consistent pattern in each of the communities that only occurred following the introduction of the intervention. That kind of thing is easy to see; it can be a mess if you have inconsistent effects.

Let's talk a little about time series. Time series designs are often used to evaluate a policy change, often within a single state or within a single governing area. They require repeated and reliable measurements, the standard methods may require as many as 50 observations before introduction of the intervention, and another 50 after the intervention is introduced. This approach relies on a combination of logic and statistical evidence. The standard methods provide evidence for a change within the group, so if I'm looking at the effect of changing the age of sale laws for tobacco products in a state, I can get a valid statistical test for whether the level is different after the intervention than it was before the intervention.

But one group designs provide no evidence based on a between group comparison, because there is no comparison group. If we include a comparison state or site and collect the same kind of data, we still don't have power for a valid analysis between the two sites because we've only got one site per condition. So, these designs have their own issues. They're certainly best used if you have an archival data collection system. They can provide good data, especially if you have multiple cycles, but if you don't have any sort of reference population you can be challenged.

Let's talk about quasi-experiments. Quasi-experiments have all the features of experiments except randomization. They also have all the problems of experiments; they just don't have the benefit of randomization. So, causal inference in a quasi-experiment requires elimination of plausible alternative explanations for the pattern that we observe in the data, alternatives to the intervention itself. If groups are assigned and members are observed, the analysis and power issues are the same in a quasi-experimental design as they as in a group randomized trial. So there is absolutely no advantage in terms of analysis requirements or sample size requirements to doing a quasi-experiment compared to doing a group randomized trial. And, you don't get the benefit of randomization, so in many cases, you're more challenged.

So why would we do a quasi-experiment? Well, sometimes randomization just isn't possible. And in that case a quasi-experiment may be a reasonable alternative; they can certainly provide experience with recruitment, with measurement, with intervention. If we analyze them carefully, they can provide evidence of treatment effects. Well designed and well analyzed quasi-experiments that were usually more difficult and more expensive to conduct than a well-designed and analyzed group randomized trial. So, I would caution anyone in the audience who thinks, "Oh, a quasi-experiment lets me do a much smaller and less expensive study." That's just not true. And if you do a much smaller and less expensive study, you're not going to have power for valid analysis.

Let's talk about step wedge designs. We mentioned these briefly in an earlier segment when we were looking at examples. These are sometimes called dynamic weightless designs, but the step wedge label has certainly caught on with most of the articles that I see published, and so this is the label that we'll use. A step wedge design combines some of the features of a multiple baseline design, with the features of a group randomized trial. And so, there's a lot to offer, with step wedge designs. Like multiple baseline designs, measurement is frequent and on the same schedule in all groups, time is divided into intervals, groups are selected at random to have the intervention introduced. So with the beginning, everyone is providing control data, control observations, and then at some point some of the groups -- or perhaps just one -- is randomly selected to receive the intervention. The other groups continue providing control data. And over time, each group receives the intervention. By the end of the study, all the groups have the intervention. Both Trials and the Journal of Clinic Epidemiology recently published whole issuers focused on the design and analysis of step wedge designs. So, I referred you to those, also to a paper by

Hughues, Granston, and Heagerty published in Contemporary Clinical Trials in 2015.

This is a diagram of a typical step wedge design, so the steps that you see are where one of more groups move from providing control observations to just providing intervention observations. In this particular diagram, the first three time periods provide control observations from all of the groups, and then group one gets moved into the intervention condition and starts providing intervention data. We move sequentially through the other groups, hopefully in a random order, and they should also start providing intervention data. At the end of the study, all of the groups are receiving the intervention.

The analysis estimates a weighted average intervention affect across the intervals, using both within group data and between group data. The approach is best used if the intervention effect occurs rapidly and is consistent and persistent and lasts. It's not very sensitive to intervention effects that develop gradually or fade over time, and I will note that these designs can be more efficient. That's one of their advantages. They may require 20 groups instead of 22, as a traditional pre-post group randomized trial might need, but they're going to take longer to conduct. If we go back to the diagram, if these time measurement periods are several months apart, it can take a long time to run through the entire study, and you're collecting an awful lot of data. In a pre-post group randomized trial, there are only two measurement occasions, you're collecting far less data, and the study may be over much faster than the step wedge design, at which point, you can give the intervention to the control arm. So, we talk about some of these issues in a paper we published in 2011.

Let's talk next about regression discontinuity designs. This is a terrific design that isn't used nearly enough in public health and medicine. Will Shadish and Tom Cook describe this design in their 2002 book. In this design, individuals are assigned to conditions based on a quantitative score that may reflect the need for an intervention. So there's no random assignment here. This is not a experiment with randomization. And it's an example, perhaps, of having your cake and getting to eat it too because it's giving the intervention to the people that most need it. That's certainly possible. The analysis then models the relationship between this assignment score and the outcome. And if you do that correctly, you can have a valid estimate of an intervention effect. The difference in the intercepts at the cutoff is the intervention effect. And several recent papers have focused on using these designs in public health and medicine. And here are references for those.

This is a figure showing hypothetical results from a regression discontinuity experiment. The upper -- or regression continuity design, I actually prefer not to say experiment even though it says that in this figure. In the upper panel, we have a situation where there is no treatment effect. And the plot is showing the relationship between the assignment variable on the x-axis and the outcome on the y-axis, and what you see is a nice linear relationship. It might not necessarily be linear, but in this case it is. And there is no change that occurs when we cross the cut point. The people on the left side got the treatment; the people on the right side got the control, but it doesn't look like the treatment had any effect because their values are not any different from what we might expect if we extended the control regression line to the left, into the lower range of the assigned variable scores. No evidence of a treatment effect.

In the lower panel, we have a different situation. It's the same pattern on the right side where we have control observations, but now on the left side, all of the values for the participants that received the treatment are elevated. The slope is the same, but there is a shift in intercepts and the entire set of scores is higher, on the outcome variable. So that's a classic example of no intervention effect with regression discontinuity in the upper panel, and a strong intervention effect in the lower panel.

If we'd done the analysis correctly, then the assignment process is fully explained by the assignment variable that was included. And if we model that correctly, we can have strong causal inference from this design. That from no less of an authority than Don Rubin, in a paper published in 1977. Regression discontinuity designs avoid randomization, but they can be as valid as a randomized clinical trial or a group randomized trial, so I recommend them. There's always a cost or a negative item associated with any design, and for regression discontinuity designs, it's power. You need more than twice as many groups in a group version of the regression discontinuity, more than twice as many people in the individual version of the regression discontinuity, than you doing the randomized trials. So, if randomization is not possible and you can employ regression discontinuity, I do recommend it. Just understand that you'll need more participants for this kind of design.Mike Pannelle [spelled phonetically] published an article in 2011, showing how to adapt the regression discontinuity design for the group context, and you can find more details there on analysis and sample size calculations.

So, the group randomized trial remains the best comparative design available. Whenever you have an investigator -- whenever the investigator wants to evaluate an intervention that operates at a group level, manipulates the social and physical environment, or can't be delivered to individuals. So if you have this kind of intervention, group randomized trial is what you should be thinking about. They provide better or equal quality evidence, and are either more efficient or take less time than the alternatives. Even so, they are more challenging. I don't deny that. They are more challenging than the usual randomized clinical trial. Individually randomized group treatment trails present many of the same issues, and investigators who are new to these designs should collaborate with more experienced colleagues, especially experienced biostatisticians.

Many alternatives have been proposed; we talked about each of these at least briefly today. Under the right conditions, these alternatives can provide good evidence. Some rely more on logic than statistics, like multiple baseline and time series. Others require studies as large or larger than group randomized trials and may take longer to complete.

Like quasi-experiments, stepped wedge designs, regression discontinuity designs.

So, I thank you for your attention today, to our discussion of alternative designs used for the evaluation of multi-level interventions. This is the last module in our seven-part course on pragmatic and group randomized trials in public health and medicine. We encourage you to visit our website and provide feedback on this module or any of the other modules in the series. You can download the slides for today, the complete reference list, and suggested activities to follow-up today's presentation. Certainly, you can view any of the modules in the series as many times as you'd like. If you have questions about group randomized or individually randomized group treatment trials, please send them to grt@mail.nih.gov, and we'll respond as soon as we can. Thanks very much for your interest.

[end of transcript]