David Murray:
Hello, my name is David Murray. I am the NIH Associate Director for Prevention and Director of the Office of Disease Prevention. I want to welcome you to Part 4, Power and Sample Size in our course on Pragmatic and Group Randomized Trials in public health and medicine. This is a free seven part, self- paced online course presented by NIH. We provide all of the slides for each of the modules that you'll see. We provide a complete set of readings and we provide guided activities for each of the modules in the course.

Today, we're going to cover Part 4, Power and Sample Size. The target audience for this course includes faculty, post-doctoral fellows and graduate students who are interested in learning more about the design and analysis of group randomized trials. We're also interested in reaching program directors, program officers and scientific review officers here at NIH who need to learn more about his kind of design.

Participants should be familiar with the design and analysis of individually randomized trials and with the concepts of internal and statistical validity. We also want participants to be familiar with linear regression, analysis if variance and covariance and logistic regression.

The learning objectives are identified here. We expect participants to be able to discuss the distinguishing characteristics of group randomized trials and individually randomized group treatment trials. Particularly, how they differ from individually randomized trials. We expect participants to be able to discuss their appropriate uses in public health and medicine and for group randomized and individually randomized trials to discuss the major threats to internal validity to statistical validity and the strengths and weaknesses of various design alternatives and analytical alternatives and to perform sample size calculations, at least for a simple group randomized trial. And in fact, that's what we're going to focus on today. We expect participants to be able to talk about the advantages and disadvantages of alternatives to group randomized trials for the evaluation of multilevel interventions.

The organization of the course is shown here. We are on Part 4, Power and Sample Size. Now power for group randomized trials is a difficult and complicated issue and we could spend days or weeks talking about it. So, the first thing that I'm going to do is refer you to some good sources of information. Quite generally we have to adapt the usual methods for power calculations that are used in randomized clinical trials to reflect the nested design that's used in a group randomized trial. One source of information is Chapter 9 in my book published in 1998. We have other text though that have been published subsequently, including the book by Allan Donner and Neil Clark. Books by Hayes and Moulton, Campbell and Walters and a very recent book that's just been out a short time by Moerbeek and Teerenstra that focuses on power for multi-level analysis. There are also some recent review articles and I've cited two of them here. I recommend all of these sources to you and the references are shown here and available to you in the reference list that you can download from our website. All of these sources provide extensive discussions on power and sample size calculations for group randomized trials.

Power is tricky. Investigators are advised to get help form a biostatistician who's familiar with these issues. If the investigators themselves are not biostatisticians, I do not recommend trying to go this alone at least in terms of power and sample size. Power for individually randomized group treatment trials is even trickier. The literature is much more limited. I offer a few references here. The work by Sherry Powell was published in 2008. I had a more recent article that came out in 2014. So, I strongly recommend that you reference these materials or send your methods expert to these materials to review them if you're interested in calculating power and samples size for individually randomized group treatment trials.

There are two penalties that we have to worry about in group randomized trials and individually randomized group treatment trials. And, this is what sets power and sample size calculations for these designs apart from power and sample size for individually randomized trials. There is extra variation. That's the first penalty. Because of the intraclass correlation, which reduces the correlation within groups, we have the group component of variation that increases the variability in condition level statistics, in group level statistics and that is the extra variation that we're referring to. If we don't deal with it adequately we're going to have less power for a proper analysis other factors constant. The other penalty is the limited degrees of freedom that are often apparent in a group randomized trial or individually randomized group treatment trial. Remember that the degrees of freedom for the analysis of the intervention affect is based on the number of groups and in a group randomized trial that may be quite limited in an individually randomized group treatment trial it may also be limited and if we don't deal with it we're going to have reduced power others factors constant if we do a proper analysis. The paper by Jerry Cornfield published in AJE in 1978 is a classic paper in group randomized trials. It's very short so I recommend it to you.

There are some strategies we can take to try to reduce the extra variation. One approach is to take random sampling within groups rather than subgroup sampling. So, if we've randomly assigned schools, it's far better to take a random sample of the students in the school rather than sampling or collecting data just in a few classrooms. If we collect data just in a few classrooms which are examples of subgroups, we're actually increasing the intraclass correlation. Where if we sample across the entire school we're reducing the intraclass correlation. So random sampling within groups is a good idea.

Sometimes we can benefit by timing the measurement in particular ways. In school studies we've seen that for many measures, the intraclass correlation is lower in the spring than it is in the fall. We don't fully understand why, but this is often the case and so I have generally tried to collect data in the spring rather than in the fall. Another thing that we've noticed in school studies is that, and this would be true in other concept studies as well, if we have a deep enough variable where there is high, within day intraclass correlation, so think about dietary intake measured in a school study. Kids tend to eat the same lunch at school, especially if many of them are participating in the free and reduced school lunch program and that's going to drive the intraclass correlation up if we collect all of our data in the same school in the same day. The way to deal with this is to spread the data collection out over time. So, go back to the school on several different days spread over a one or two- week period and that will drive the intraclass correlation down compared to what it would have been. And here are references for some of these issues.

Other reflective strategies and this is one that I strongly recommend to you, regression adjustment for covariance. Now this is one that we deal with at the analysis stage rather than the data collection planning stage, but it's still very useful. Intraclass correlation is an example of confounding of sorts. We have characteristics associated with the groups. If we can measure those then we can adjust for them in the analysis, these would be fixed covariant in non-repeated measures analysis. Time veering covariants and repeated measures analysis. And this is one of the most effective ways to make intraclass correlation smaller. It also serves to reduce residual errors so that's another benefit. Often we can reduce the ICC's by 50 or 75 percent with a dramatic improvement in power. So, I strongly recommend this approach to anyone who's considering analysis of data from group randomized trials.

These are some strategies that I do not recommend. They have been suggested at various times. Used at various times, but they have been shown to be not only ineffective but actually dangerous. So, if we use individual degrees of freedom, we're going to have an inflated type 1 error rate. If we use kitchens effective degrees of freedom, we're going to have an inflated type 1 error rate. If we base degrees of freedom on subgroups, rather than the groups that we've randomized, we're going to have an inflated type 1 error rate. Those are better known. The fourth one on this list is less well known. If I analyze my data with a repeated measures mixed model analysis of variance or covariance, and I'm modeling more than two time points in my analysis, I'm going to have an inflated type 1 error rate. It increases the degrees of freedom, but it results in an inflated type 1 error rate. So, I don't recommend any of those.

The best way to increase degrees of freedom is simply have more groups. More members doesn't help so much though it will in an individually randomized group treatment trial if I only have small groups in one arm. But in group randomized trials it's the number of groups that matters and the best way to increase degrees of freedom is to have more groups.

Sample size, detectable difference in power. So, in the book that I wrote in 1998, I talked about seven steps in any power analysis and I think this is generally the case, whether it's an individually randomized trial, like a group treatment trial, a group randomized trial. And the seven steps are listed here. We're going to walk through each of these but they apply quite broadly to do in sample size calculation in a variety of context.

Intervention affects in my work I try to define them as one degree of freedom contrast. If I do that then the t-test is the appropriate test. It's very easy. We know what the distribution of the t-test is. Critical values are easily obtained. Once I settle on a type 1 and type 2 error rate and so that helps tremendously. In my book and in other sources that you can find we provide formulas for the variance of the intervention affect given a variety of design and analytic plans which brings us to the sixth step which is gathering estimates of the perimeters that define the variance of the intervention affect.

I want to make a point here, that those estimates are best collected from data that are similar to the data that we're going to collect in our study. So we want to get estimates from a similar population. Data collected using the same measures. If possible, data collected using the same or similar design and certainly estimates that are based on a similar analysis. To the extent that we differ and get our estimates from a different population using different measures based on a different design, a different analytic plan, the estimates may be no good for planning the study that we're going to do and it's just guess work. So, it's really important to get good estimates and get estimates that reflect the kind of study that we're about to do.

How do we estimate an intraclass correlation? Well, one way is to go to the literature. Increasingly we can find intraclass correlations published in the literature. The consort guidelines now ask that investigators publish the intraclass correlations and that's helped the situation. So, we can often go to the literature and get good estimates. Another way is to estimate the intraclass correlation from your own data and if you have pilot data from the study that you're getting ready to do that can be a good source. Simply apply a one-way analysis of variance with group as the only fixed affect and apply this formula and you've got an estimate of the intraclass correlation. If you want to try to shrink that estimate, you can adjust for covariance in this analysis and now you have a one way and [unintelligible] with group as not the only fixed affect, but one of several and now again you're taking the MS between and within and calculating an adjusted intraclass correlation.

The seventh step is to actually calculate sample size or detectible difference or power based on the estimates that you've gathered. Step number, I'm actually going to back up a couple of slides and look at my original list. The fifth step there, developing an expression, can be tricky, but my book is an excellent source and there are other places where you can find expressions of the variants. Gathering estimates of the perimeters is perhaps the most difficult step. Very important, we talked about doing that for a moment. And then, actually calculating the sample size detectable difference or power is fairly straightforward once you've gone to the work to figure out what is the variance of the intervention affect and what are the estimates that I can apply? In doing this we use formulas. If I have defined my intervention affect as a 1 degree of freedom contrast, for example the intervention mean compared to the control mean or the intervention mean slope compared to the control mean slope. The detectable difference in a simple randomized clinical trial is shown here. And the second expression which is equivalent to the first is the version that you're probably used to seeing. We have the usual variance of the dependent variable, sigma squared y, divide by the number of observations that contribute to each mean in. We multiply by the sum of the critical t values reflecting the type 1 and 2 error rates. We square that sum. We multiply by 2 because we're comparing two means and then we take the square root of all of that and we have a detectible difference. So, it's a very straightforward calculation. It's not iterative because nothing's going to change. We may decide that that detectible difference isn't, isn't adequate and so change the sample size for the study, but otherwise it's not an iterative activity.

In a group randomized trial, it's a more complicated expression. At least the second one, the first one you'll notice is exactly the same. The variance of the intervention affect multiplied by the critical values by t, summed together and squared. The lower expression adds the parenthetical expression that we saw in the first part of this course, module 1. The variance inflation factor or design affect defined by the intraclass correlation and the number of observations that were measured in each group and are contributing to the analysis. The dominator instead of being n, is now mg, where we have n members per group and g groups per condition, we still have a 2 as a multiplier because we're comparing two means or two slopes. And we have our t squared values, or the t values, the sum is squared off to the right. I always use the t distribution, never use the z distribution because the z distribution assumes an infinite number of degrees of freedom. That never applies in group randomized trials and so I use the t distribution and caution you to use it as well. If you use the z distribution, you are not taking the penalty associated with the limited degrees of freedom into account and you need to do that.

This slide, and the next two that follow will help you understand a bit the role of the intraclass correlation. The number of groups per condition and the number of members per group in the detectible difference calculation. You've seen these before if you watched Part 1 in this course where we reviewed them briefly. The lines show the detectible difference which is the, on the y axis. The lines distinguish based on the number of groups per condition. The red line that's at the top has only two groups per condition. So, that's the smallest group randomized trial that we can run and have any variation associated with the group within study condition.

Members per group are shown on the x axis and you'll notice in general that the lines have a little bit of curvature to them as we move from 25 members per group up towards 100 members per group. After that the lines are pretty flat. And what that's telling us is that when we have a large intraclass correlation like this, .1, we don't get much benefit out of having more than 100 or 150 members per group.

In contrast we get a lot of benefit out of having more groups per condition. The line with just two, that upper red line is quite high on the y axis and the lines drop as we add more and more groups per condition. This slide also shows that if you only have two or four groups per condition and an intraclass correlation is high as .1 you have to have a very large intervention affect to have, to find it to be significant because the detectable differences are quite large in standard deviation units. And, you have to have a quite large study in order to be able to detect a more reasonable intervention affect. As the intraclass correlation shrinks and here it's dropped by an order of magnitude, we get more curvature in the line that gives us more benefit from adding members to each group. All of the lines drop and now we can have a reasonable detectible difference. Even with 16 or eight or sometimes fewer than eight groups per condition.

If the intraclass correlation drops by another order of magnitude then lines drop a little further, there's more curvature and we are usually in pretty good shape, even for smaller studies. It's very important that you do your own calculation based on a good estimate of the intraclass correlation. It's dangerous to assume that the intraclass correlation is going to be as small as this or even as small as .01. Much better to have a data based estimate for the study that you're proposing.

Now, let's actually walk through an example, particularly an example using the seventh step, which is calculating the sample size. If we have a one degree of freedom contrast between two means or two means slopes, the sample size calculation for the number of members in a randomized clinical trial is the first expression and the corresponding expression for a group randomized trial is at the bottom of the screen. I showed you the formula for detectible difference earlier. This is the formula for the number of groups we need in each condition. It involves the same elements algebraically, but they're rearranged to solve for g.

Let's calculate the required sample size per condition for a two conditioned randomized clinic trial first. With a 5 percent type 1 error rate, 80 percent power, you've seen the formula that's used. It's the standard formula. To make things simpler I'm going to work in terms of standard deviation units and set sigma squared y to 1. If we substitute this expression into the formula and use a delta of .2 if we assume that we have lots of degrees of freedom, I can use 1.96 and 0.84 as my critical values for t. Do the math and we get a number of 3.9, 392, sorry, for the number of members per group. So, we need a pretty large study with a modest affect size of .2 standard deviation units, but here's the answer and we're done. If we want to apply this in the context of a group randomized trial one of the things that we're going to discover is that it is an iterative calculation. And, that's because the number of members per group determines the critical values and the number of members per group appears in the numerator and the denominator and the value that we get for g may change from one calculation to the next as we move through the iterations.

Here we're going to make the same assumption in terms of the type 1 and 2 error rate of 5 percent, 80 percent for power. The same magnitude intervention affect, .2 standard deviations. We're going to start with, because we have to start somewhere, an intraclass correlation of .01 and 100 members per group. If we drop those values into the formula and do the math, the value that we get for a result is 7.8. We always round up so eight groups per condition. Now, we can't stop there though, eight groups per condition because the t values that we used in this expression assumed an infinite number of groups per condition. So, that the critical value is at 1.96 and 0.84 don't hold if we only have eight. If we have eight, we can recalculate degrees of freedom, 2 times g minus 1 or 2 times eight minus 1, 14. Now, we need to go back to our t table, look up the critical values based on 14 degrees of freedom and drop those into our formula, 2.145, 0.868, redo the math, we get 9.03. So, now we can drop 9 into our formula for degrees of freedom, we get 16. We look up the critical values of t there, 2.12 and 865, drop those in, do the math, we get 8.86. We can round up, we're at nine again. At this point we can stop because the value that we're getting out of this calculation, nine is the same value that we used to calculate the degrees of freedom and the critical values for t. So, now the critical values using the expression are consistent with or concordant with the result that we're getting and we can stop. We can say that if we randomly assign nine groups to each condition, we will have given a type 1 error rate of five percent, 80 percent power to detect a .2 standard deviation affect given an intraclass correlation of .01 and 100 observations per group included in the analysis. I said we could stop there and certainly we can stop there for that calculation. We don't have to do any more iterations. But I would caution you it would be very wise to perform a sensitivity analysis using different values of the intraclass correlations, different values of m, different values of g to see how much the result would change as these other factors change and you may want to present some of that in your grant proposal. I always do.

Let me comment a little bit on the situation that we deal with if we have an unbalanced design. All of the formulas that I just showed you in the example that we worked through assumed that every group had the same number of observations contributing to the analysis. And as long as the ratio of the largest to the smallest isn't any worse than about two to one, okay, we can use those methods. If there is more extreme imbalance though, other methods are required. You can have an inflated type 1 error rate if you use the, what I call the standard group randomized trial methods if the imbalance is worse than two to one. And I'm going to, not walk you through the methods but I am sharing with you here references to recent articles that address this issue and I recommend that you take a look at them.

If there, if you're working with an individually randomized group treatment trial and have substantial imbalance in the group size, that's also been addressed recently in the literature and here are a couple of references. We can get into difficulty here as well if the imbalance is worse than two to one.

So, in summary the usual methods for detectable difference, sample size and power are not appropriate for group randomized trials and individually randomized group treatment trials. We have to adapt those methods to reflect the nest of designs that are inherent in these other studies. Power is a tricky subject and I strongly recommend that investigators who are not biostatisticians themselves work with a biostatistician, 30 years ago it was difficult to find a biostatistician who was familiar with these methods. That is not true today. Every decent sized campus is going to have biostatisticians who are familiar with mixed modeled regression, random affects models and even if they haven't heard of a group randomized trial or a cluster randomized trial, they will be familiar with the underlying methods and can do a little reading and come up to speed on it fairly quickly.

You need to address both of Jerry Cornfield's penalties, extra variation and limited degrees of freedom. It's increasingly common for people to remember the first part, but they often forget the second part and I want to emphasize both. Often we'll see in the write up of a paper that they used a method to account for the extra variation but they report a z test or a wall test that assumes infinite degrees of freedom. You need to deal with the limited degrees of freedom as well and be reporting a t or an f or something that reflects the limited degrees of freedom. Failure to do so will result in an inflated type 1 error rate and none of us want that. There are effective design and analytic methods that we can use to reduce the intraclass correlation. Regression adjustment for covariant is one of the best, but there are also some design steps that I mentioned.

The most important factors affecting power in a group randomized trial or individually randomized group treatment trial are the intraclass correlation, the number of groups per condition and so you want to have a good estimate of the intraclass correlation and have a good number of groups per condition in order to have adequate power.

I want to thank you for your attention today. We've been talking about sample size and power in a group randomized trials in public health and medicine. I want to draw your attention again to our website where you can give us feedback on this module. You can also download the slides from today's presentation. The references for the entire course and you can download suggested activities based on today's presentation. You can also watch the module again if you'd like to see it and review the material. I would also encourage you to watch the next module in the series which present some examples of group randomized trials. If you have any questions, please send them to grt@mail.nih.gov and we'll get back to you. Thanks very much.

[end of transcript]