ICF Transcription
ORE Webinar Power Analysis Part 3

LILY: Greetings. Welcome, and thank you for your interest in this power analysis training session. My name is Lily Zandniapour. I'm a research and evaluation manager at AmeriCorps, and, on behalf of the Office of Research and Evaluation at AmeriCorps, I am pleased that we can provide you with this third training in our three-level series of training modules on Power Analysis for Program Evaluation.

The first-level training webinar provided an introduction to power analysis. The second training module covered the basic mechanics of power analysis. This third-level training module's focus is on applied power analysis. This training series is delivered by our evaluation training and technical assistance provider, NORC at the University of Chicago. For some years now, we have partnered with NORC to provide support to our grantees and strengthen their evaluation studies so they could use credible and quality evidence about the programs and interventions they implement and that the agency supports. We hope that you find this training, as

well as the other two modules on this topic, helpful in your work.

Dr. Carrie Markovitz, who is the principal research scientist at NORC and project director for our evaluation training and technical assistance work, will lead this training and introduce our presenter, Dr. Eric Hedberg. Thank you.

CARRIE: NORC at the University of Chicago has been collaborating with AmeriCorps' Office of Research and Evaluation for almost a decade to strengthen the existing evaluation guidance and tools for AmeriCorps state and national applicants and grantees.

Over the years, our TA team has assisted and supported numerous AmeriCorps grantees with their evaluation plans, design and implementation challenges, instrument development and reporting. And through this work, we have identified common areas of need and then worked with the Office of Research and Evaluation on developing tools, webinars, classes and presentations to fill these gaps.

So this class on power is one of several classes on evaluation topics that were developed over the past 10 years by N-O-R-C, or NORC, and are currently available online on AmeriCorps' website. These classes include information on topics from logic model development to drafting research questions and addressing other design topics to budgeting and managing and evaluation. We encourage everyone to check out these other classes based on their own evaluation needs.

Now, I would like to introduce Dr. Eric Hedberg. For Dr. Hedberg, power is a major methodological topic of interest. Most known for his work on evaluation design, Dr. Hedberg is an interdisciplinary quantitative methodologist. His research interests include several areas of methodology related to evaluation and analysis, and he recently authored a sage [phonetic] little green book on statistical power analysis.

Dr. Hedberg is an accredited professional statistician by the American Statistical Association, and he is a sociologist. His current areas of research include investigating the design of evaluations in education and criminology in addition to measuring social capital through social network contextual effects. Dr. Hedberg has authored or co-authored over 30 methodology-focused papers and books that have appeared in education, medical and criminological journals, while also contributing to numerous reports and presentations at major research conferences. Dr. Hedberg earned his PhD in Sociology from the University of Chicago.

So now I would like to pass this to Dr. Hedberg to do his presentation.

ERIC: Thank you, Carrie, for that great introduction. So hello, everyone. Welcome to Level 3 of Applied Power Analysis, presented to you by AmeriCorps. This is part of a series on Power Analysis for Program Evaluation.

This is the final deck in a sort of three-deck series. The first level, Program Analysis for Power Evaluation, really sort of tried to sell you on the idea that power analysis was an important thing to do. After that was the second deck, Level 2, which just sort of provided the basic mechanics of power analysis – really didn't go into a lot of the technical details, per se, but hopefully gave you some intuition as to what aspects of study design and [inaudible] go into the chance that you'll have a statistically significant result.

Now, in this deck, Level 3: Applied Power Analysis, we're going to get into a little bit more technical detail and actually even do a couple power analyses live using some freely available software. So this Level 3 is intended for program staff and evaluators, and specifically those working with statisticians. Some of you may be doing power analyses yourself. Others may be doing power analyses in concert with consultants or university faculty or people from places like NORC, statisticians – and so this is a

useful deck, I think, to help you have conversations with those folks helping you.

So we're going to talk about how to conduct a power analysis – what parameters and what values you sort of need to bring to the table to do your power analysis – we're going to talk a little bit about estimating effect sizes for trying to read the future and sort of get a sense of what you think your program impacts might be and so how to power for that, and then we're going to talk about some examples of common designs we often see in evaluations. Finally, we'll end with a bit of a discussion, and I'll highlight a handout that'll be available on the website that sort of gives some examples, and we'll talk about how to write up or even read a power analysis.

Just sort of situating this deck in the larger ecosystem, Level 1, again, was just sort of getting an idea what statistical power was, introduced the concept and hopefully sold you on why it was important. Level 2 was about the basic mechanics,

what kind of thing goes into it and what are the sort of major swings that tend to go into power analysis – namely the Type 1 and Type 2 errors.

Again, just so we're all on the same page in case you're watching these webinars or these decks weeks apart, on our last episode, we talked about, what is power? And again, power is a probability – the probability that you'll get a statistically significant result. And what this is is, this is a result that's based on many things. It's based on the design of your study, namely how your groups are formed, how you picked observations – did you pick everybody, or do you have a sample? How did you analyze the data? And then again, it's also dependent on the amount of data that you have, how you analyze the data and the size of the impact as it exists of the thing that you're trying to detect and what significance level you're comfortable with.

Just to reorient you into where statistics sit in the larger logic of study, again, we have larger populations from which we often draw samples. Those

larger populations have parameters that we typically want to know, like the impact of a program or average income or total number of hospital beds or something. And from our small samples, we tend to estimate statistics, which are best guesses as to those larger parameters, which unfortunately we never actually see. And so the idea of powering a study adequately is so that we can trust the statistic we estimate as a good inference as to the likely parameter.

We also talked about, in the last deck, the Type 1 and Type 2 errors. We talked about how there are four results of studies that are possible based on two ways to be right and two ways to be wrong. We could be right in that, when there is not an impact, we could say, yep, there's no impact. That could be correct. Or if there's an impact, we could also say there is. And we could also be wrong in two ways where we could say, ah, but – if there is no impact, but we do say there is, that's Type 1 error, and finally we could be wrong as that there is an impact, but we conclude that there is not an impact. That's

Type 2 error. And that's really what the main drivers of power are.

We talked about, briefly, Type 1 error and that typically we want to set that to be about a 5-percent chance of Type 1 error. That's typical. However, changing that threshold will have an impact on power. This is a risk-reward tradeoff. And so we generally want to set Type 1 error to the conventions of the discipline, 0.05 or below, and then we want to set power to be high enough that we have a good chance of detecting the effect, should it exist.

We talked a little bit about how all of this relates to these curves. [Inaudible] a second here to really dive into these curves a little bit. The first curve in the solid black line – this is our Type 1 error. This is the idea of the results of all the possible samples if, in fact, there is no impact, and the idea is that this is our sort of worldview when we do a null hypothesis test. And what we want to do is, we want to say, okay, if our data is in those black-shaded tails, if our data is so unlikely given this

worldview that there's no impact, we may want to reject that idea and say there likely is an impact.

And so power analysis is about drawing the second curve, the one that's sort of purple and blue. That's the representation of all the possible samples that we could get from a world in which our program does have an impact. Now, it's a range of samples. Some of those samples will not be significantly significant. Those are the light blue samples. If we pick one of those, we won't be able to conclude that there's an impact.

However, if we pick one from, say, the purple part of that second curve, those are samples that would produce a statistically significant result. We don't know which of those samples we're going to get, and so power analysis is all about understanding and creating a large enough sample that it moves that curve all the way we can possibly can to the right so that, while we don't know what sample we're going to get, there's a really good chance, a really good

probability that we'll take a sample that gives us a statistically significant result.

Again, why do we want to conduct a power analysis? This, I think, is an important thing, not only in its own right, but power analysis encourages you and your research team to really think critically about all aspects of your study. Doing a power analysis check at every decision point keeps everybody honest. It really makes you think hard about what your expected impacts are. It makes you think hard about the design of your study, makes you think hard about your analysis plan, and all of this is connected to the budget.

And so as you make decisions as to how big your sample needs to be, should you do fixed or random effects – which we'll talk about in a little bit – should you randomize, say, within hospitals or between hospitals – all those decisions have implications, not only for your budget, but also have implications on the power of your study as well. And so, as you make those decisions, work with your

statisticians or work with some of these tools we're going to show here to just do a gut check, make sure that you're maintaining the power of your study as you make those decisions.

So power analysis is not a one-and-done thing. It's not a check box as you go from planning your study to carrying it out. It's really something that you want to evaluate and think deeply about throughout the entire planning process. And I really say planning process because power analysis is very important to be conducted before any data are collected. Once you've collected your data, we're no longer talking about this curve, this range, this set of possible outcomes. You have your data set. You've picked your one sample.

And so a power analysis based on that isn't as meaningful as some often think of it. And there's a little site below that you're welcome to google and try to read about and get sort of that understanding about this. So again, before data collection is really the time to be doing a power analysis.

How do we design a high-powered study? Well, what I like to do is, I like to break down a lot of these complicated statistics and things and the real key components that go into it. So I'm going to break it down myself into the study design, the statistical analysis and the expected impact. And we can think of this as sort of the slow chart here of, the study design really sort of drives our choices, and our statistical analysis attempts to detect the impacts that we expect. And all of that goes together to produce the final result of our power analysis.

So let's start breaking this up a little bit. Let's sort of really dive into this. Only the statistical analysis that you choose [inaudible]. Power really, then, is sort of this function of the study design and the expected impact. Think of the study design in the sort of old physics experiments or even new physics experiments, where you're really trying to build some mechanism to detect something. This could be as simple as a filter and trying to pick up lightwaves in some kind of physics experiment, or it

could be a very expensive machine like the particle collider that's taking up half of underground Europe. All Of these things are basically machines that are being designed to detect an impact.

Well, your sample and the surveys or the administrative data that you analyze are essentially a machine built to detect an impact of your program. And so the power of your particular study really is an interplay between the machine you built, the study design and the impact that you can expect.

And so power analysis often breaks down into two questions that are related but sort of trying to figure out two aspects. How much data do you need to detect an impact? Or, if data, the size of your sample is something that you really can't control and you're sort of stuck with the sample you have, then you want to ask the question, okay, what's the smallest impact that I can detect with this particular sample? And then of course the statistical analysis design links these two to ultimately produce power.

So we sort of want to think about this triangle of, we can organize all of our questions about power as sort of, what does the third thing need to be holding the other two constant? If we know the effect size and we know the sample size, what's our power? Or, if we know what we want our power to be, and we know the effect size, how large does the sample size need to be? Or, if we know what the power's going to be, we know what our sample's going to be, we can ask, what's the smallest effect size we can possibly detect?

In very broad strokes, of course, larger effects tend to have more power, holding other things constant. Larger samples, more or less, typically have more power than the smaller samples. So the composition of the sample sometimes makes a difference, which we'll talk about in a little bit more detail in a couple of slides.

So for the last 10 minutes, we talked about a lot of the things that were mentioned in the other decks,

Level 1 and Level 2. Now we're going to slow down and start getting into a lot more of the details. Over the next few slides, we're going to talk about how the type of outcome, linear continuous outcomes versus categorical – how those have an impact on your power analysis. Then we're going to dive deeper into the study design – how the groups' treatment and comparison are formed, aspects of sampling, any other measures that you may have and then finally sort of the choices we make in the analysis plan. All of these aspects of study design are going to have a major influence of power.

So let's start talking about the types of outcomes. First we're going to talk about the linear outcomes. Linear outcomes are things like money, test scores, your, say, BMI – these are numbers that have meaningful differences between them. Sometimes they're skewed. Say, like, income – it can be transformed to get back to more that sort of bell shape that we like. Other types of outcomes are sort of these unordered categories. They have a limited number of possible values, and they're not really

ordered. The most basic example would be sort of a yes-or-no variable. Say, did you get a job? Yes or no? Did you graduate college? Yes or no? They tend to have opposite values at or above/below grade and so on and so forth.

Often, these unordered categories can be created from linear outcomes if you so choose. Other possible unordered categories can have more than independent values. After high school, or, say, if you look at ninth graders, did you matriculate to college, or did you graduate high school? Or did you drop out? These are sort of unordered categories with three distinct outcomes for any ninth grader.

We could also have [inaudible] ordered categories. We could sort of have the analysis of ordered categories sometimes similar to linear or categorical. A lot of psychology papers tend to do this. Often, these are shown as satisfaction questions. How strongly do you agree or disagree with a particular policy, or how satisfied are you with your position? Often, you are given choices

that range from, say, one for strongly disagree to five, strongly agree, and you give your best answer.

But the difference between, say, a three and a four can sometimes be nebulous. Sometimes people want to analyze these as linear outcomes. Other times, people want to analyze them using more complex methods. Again, if we treat them as categorical variables, one important thing to know – that you typically need more data to analyze categorical outcomes than you need to analyze linear outcomes.

Next, we're going to talk about study design, sort of this group formation. And this is a very important thing, because evaluations, more often than not, are focused on differences between those who receive the intervention compared to others who did not. How people are chosen to receive an intervention versus how people are chosen to not has a major impact, not only on your broader design and the inferences you can make, but it will have an influence on your power analysis as well.

Group assignment can be broken down into many different ways. The main stars of the show tend to be a random process. These are sort of the randomized controlled trial. Random assignment has the attractive feature that now your treatment and control variable is generally uncorrelated with anything else that you have in your data set and not correlated with the stuff you don't have in your data set, either.

Another way to form groups is, you could match. You could say if you have a group of people receiving the intervention, you can use some statistical method to generate a comparison group from other possible observations. Oftentimes, this is analyzed just like a random sample, but the value of covariants and the inferences you can make are a little bit different from the purely random assignment.

You could use, say, a threshold. You know, the sort of regression discontinuity designs will often say, we have a sort of forcing variable, or some people call it an assignment variable, where basically you

have a range of individuals, and people at or below a cutoff will receive a treatment versus, the rest do not, say, you know, being 200 percent over the poverty line.

And then finally, there's natural assignment, which is the least rigorous, but sometimes it's what you have, which is, you know, some treated, some not, and it's usually the result of some other process that's usually systematic and correlated with many other things, so you have to kind of keep your eye on that.

Another aspect of study design is sampling. And sampling's a process that – you have a population that you want to infer, but you only have the budget, time or ability to collect data or sort of a subset of that. There are many ways you can sample data, and there's volumes of books, entire shelves in your library, that talk about this. The two that are most applicable to evaluation design are these sort of simple samples that we often assume are simple random samples, in which everyone in the population has an

equal chance of being selected, and you analyze the subset as the simple sample.

However, there are more complex ways you could draw a sample. You could create a multilevel sample by first sampling some schools within a district. And then once you've selected your schools at random, you select either everybody or a sample of students within those schools. And so these complex samples need to be analyzed in a different way than samples that are just a simple random sample. And that has a big influence on power.

In complex samples, we want to hearken back to the [inaudible] assignment issue in that we have a complex sample, say, students in schools – you have a choice at the design phase of whether you want to control the assignment of treatment versus comparison at the entire cluster or school level, or you could do it within school. There are tradeoffs of both ways, and perhaps future webinars could maybe dive into some of those.

And then finally, in your study design, you could possibly measure other things. You could ask other questions or even collect data prior to the assignment of treatment or control. If you do collect data that are similar to your outcome measure, then you have a pre-/post- set of measures, and these can open up a wide variety of different analyses, choices [inaudible] or other things. You could measure many pretests and many posttests in which you would have sort of a time series. Or you could measure other things prior to random assignment or prior to the creation of the comparison groups that may not be the exact same metric as, say, your outcome, but are highly correlated. One example would be, state standardized tests could be used as a covariant to try to control for prior achievement or prior ability if the dependent variable is, say, the SAT, which they never had the chance to take prior to your study. And so a lot of these things can sort of go into and have an influence on the power.

What we see here is, this is sort of a map, and it's not meant to be exhaustive at all, but just to sort

of give a sense as to the wide variety of possible designs. One in the orange is this one group where it's only the intervention group, and that sort of leaves you with having to do a pre- or/and post-analysis. Now, from there, we could say, let's just measure one time point before and one time point after, in which case you have either a simple sample, or it's possible that you could have a complex sample.

Or you can have many time point before and after in which we're talking about sort of time series. When it's just a single case study, that's analogous to sort of a simple sample, or you could have complex, where you follow many people, and this is sort of what our friends in economics call panel studies. A more rigorous design would be to have two or more groups, an intervention group and a control group and maybe different versions of intervention.

From here we could have only the posttest, and we could just have a simple sample. Or we could have only a posttest and a complex sample. And from there

we can choose, do we want to assign within? Do we want to assign treatment between? If we have pre- versus post-, we see a whole plethora of options. We could have just one time point before and after, in which case we could, again, have a simple sample with a post- and a pre- control, or we could have a complex sample, which, again, leaves us deciding whether we want to assign within or between.

We could also do difference/indifference models, which would yield a complex sample, and then finally, one of the more complex designs – many time points before and after, in which we're sort of in the land of interrupted comparative time series and all sorts of options there.

The important thing to note here is that every little circle here is a different test. It's a different analysis, and it's a different test, which means all of these are different power analysis procedures. One of the biggest mistakes I often see is a great study design and then they use a power analysis that has nothing to do with what they're actually doing,

and oftentimes this is the sort of check box thinking that goes into power analysis. So it's very important to realize that all of this variety means that there's an equal set of variety in the different power analyses you could perform.

In your analysis plan, a lot of your choices in study design will then present you with some choices or some decisions as far as how you analyze your data. Your type of outcome means you have to take different procedures in your statistical software. Your group formation will decide the sort of statements you can make from your analysis. Your type of sample will also influence the sort of buttons you have to push in your statistics software, and other sorts of measures can lend themselves to nuances versus regression or just comparing groups with Enova [phonetic] or other things. Again, all of these things are different power analysis.

I want to take a quick second to talk about other measures and power in a little bit more detail. Using covariants in your analysis, especially when

you randomize, is an incredibly powerful tool to increase power, because essentially what it's doing is, because these other measures are uncorrelated with your random assignment, they will explain variation in your outcome. And your statistical [inaudible] depends a lot on the unexplained variation in your outcome. And so having other information that you can put into your statistical analysis can really help with that.

One thing you want to be careful of is, how your groups are defined will really have an impact on the usefulness of having other measures or covariants, as we say. If you have randomization, I think you're in pretty good shape, because the randomization itself rigs it so that treatment versus control is going to be not at all very correlated with your other covariants. However, if you haven't randomized, using control variables can sometimes hurt you more than it helps you.

There's a great paper by Porter and Raudenbush that talks about Encova [phonetic], and this graph is sort

of a way to think about one of the main messages that I got from that paper. The idea is that your estimated effect – you know, what your model finally shows, all the way here on the right – is really the result of the raw actual effect minus the group difference in your covariants. Say if you have a pretest of a state test score and you have treatment versus control, this group difference in comparison in your covariants – sort of, how different are your treatment and control groups on, say, their eighth-grade tests? And then how much is that correlated – that's that third circle – how much is that correlated with your outcome? If your outcome is their junior year of high school SAT score, their eighth-grade math score is probably going to be pretty correlated. Those two combine to take away from the actual effect when you actually estimate it with your model.

So what you want to be careful of is, if you do use a covariant, you want either the differences between treatment and control to be very small, and/or you want it to not be related to the outcome at all. If

it's not related to the outcome at all, it's not very helpful anyway, and so it's always important when you're doing these studies, whether it's randomization or whether it's matched, to actually check this difference in the covariant between your treatment and comparison groups. That helps ensure that you get an unbiased estimate of your actual effect.

So finally, let's put all this detail back into the big picture. So your power for your study is the result of group formation, study design, your outcome type and your analysis procedure. Your sample size is one of the biggest components of your sampling design that has an influence on power. And finally, your analysis procedure is designed to really sort of detect this effect size. And so power analysis is really trying to say, if I have a sample, what's the effect size that I can detect, or, if I expect a certain effect size, what sample do I need to get? And this is a different representation of this triangle that I've been showing you. And so this is

sort of a visual way to really understand all the different mechanics that go into power.

Throughout all of these slides, I've actually had a lot of slides that have talked about effect size, and I just want to take a second to really sort of talk about effect sizes and what they are. So effect sizes are basically expectations of the impact that your program is going to have on an outcome. There are many different types of effect sizes based on the type of analysis that you perform, but the most common is what's called Cohen's d, which is sort of the difference between groups and standard deviation units.

If your program is really trying to influence individuals and it's trying to influence the social world, social systems, as we all know, are vastly complex, and that means the effect sizes, because of all that complexity – it tends to create a lot of variation, meaning that effect sizes, which are the differences of the impact divided by the standard

deviation, tend to be smaller than we would like them to be.

This doesn't mean your program isn't effective, but is it more of a representation of the vast varieties there are in human beings. Other programs, say, environmental programs that are removing invasive species, tend to have larger effect sizes because they are creating a much more noticeable difference in a situation in which there's far less variety. [Inaudible] either have weeds or they don't. And so the effect sizes for environmental programs compared to, say, a program trying to prevent high school dropouts are going to be vastly different.

When Cohen first wrote his book on power analysis and did a lot of his work, he came out with an article that tried to give some guidance as to what effect sizes are in terms of small or big, and he said 0.2 is small, 0.5 is medium and 0.8 is big. He based a lot of this on his own set of psychological studies and his sort of survey of [inaudible] psychological field. That means it has nothing to do with

environmental impact; 0.2 might be incredibly small for an environmental study; 0.8 might be noticeable, but not helpful.

So the important thing is that, whenever you're planning your power analysis, the effect size that you choose needs to be based on the literature. It needs to be based on good assumptions. And that could be sometimes a tricky thing to do. And so what's important is, don't guess. What's important is, don't use [inaudible] sizes. What's important is, you look at the literature, and there's a couple ways you can go about doing that, which we'll discuss in a little bit.

Next we want to talk about how to perform a power analysis. And we're going to – I'm going to show you some of these tools, and we're going to walk through some examples. Computations for power analysis tend to be sort of straightforward for the more straightforward designs. If you're just comparing two groups and you're not measuring any other variables, the power analysis formula's not

horrendously scary, and it's a straightforward thing to do. Once you have covariants or a complex sample, or you're doing matching, or you have a different weight, different observations, or some people don't answer the question and you want to do non-response, this all adds a lot of complexity, and so the power analysis, again, will become more complex.

There's lots of software out there to provide estimates based on your study design and analysis plan. Most of the major stats packages, SPSS, Stata, SAS, R, Python, already have built-in libraries to handle the most common things. There's other pieces of software, PowerUP! or G*Power, that are one tool to do nothing but power analysis, but they can do power analysis for a wide variety of outcomes.

When you walk in to do a power analysis, you actually need to bring to the table some key bits of information. So what information is needed for a power analysis? First, you'll need to have a good handle on your study design, especially a lot of numbers. How many groups do you have? How many data

points do you have? What sampling method? You want to walk in with an assumption of what you want your power to be; 0.8 is a pretty good convention. You also want to walk in with what you want your significance level to be – again, 0.5 for our classic [inaudible] statistical significance test is the current convention.

Then you need to sort of understand your effect size. You have to have an estimate. What do you think your impact is going to be? And then, finally, you need to have some expectation of how much sample you can afford or how much existing program data. Once you have all this information collected, or at least some intuitions about this, you're ready to do your power analysis.

I promised before we'd talk more about estimating the effect size. As we talked about during the Q&A in the first level, estimating the effect size can often be tricky, because if somebody has already done a study on your intervention with your population with your design, then you already have a good sense as to

what your effect size is going to be. But unfortunately, replication is still relatively rare in most of the social sciences. And so it's very unlikely that you're going to be able to find a study that did exactly what you plan to do to give you advice on what you should plan for.

And so where do you go from here? Some useful strategies for doing this is, find as many studies of similar programs as yours, and based on the type of intervention or the population served and the outcomes measured. What this does is, if you look at people who are doing a similar intervention to yours, maybe with different outcomes, you can at least get a sense of how much interventions like yours tend to change other outcomes.

Another way to do it is to look at studies that say, okay, I want to look at how other interventions – maybe interventions completely different than mine – have had an impact on the outcomes that I'm trying to move. How much is the needle movable on my outcome? So if you look for studies in those two areas, you

kind of put different hands on the elephant, as it were, and try to start to get a good sense of what you could possibly expect.

Another great source is meta-analyses. Meta-analyses take lots of studies. They convert their statistical tables into this common effect-size metric and then estimate an average effect size. So meta-analyses are of the two things that we just talked about – meta-analyses of your outcome or meta-analyses of similar interventions are also useful in giving you a good sense as to what's sort of out there.

Okay, so now that we've discussed where you could possibly find effect sizes, I think now's a great time to sort of dive in and actually do some power analyses and do some examples. What we're going to do is, we're going to take one study, one hypothesis that we want to test, and we're going to do many different versions of that, variations on the study, variations on how we would analyze these things, and start to get a feel for how different power analyses can turn out and all from the same study.

So what is this study? So let's suppose that there's a school-based math intervention program for middle-school-aged youth. And it wants to determine if the program is having an impact on students' standardized math scores – say, the math scores that are provided from state standardized tests. So here's where we're going to break [inaudible] for power. We want to assume a power level of 0.8. That is, we want to assume an 80 percent chance of being able to detect an effect.

Our significance – we're going to assume a Type 1 error rate of 0.05 per convention [inaudible]. Our effect size – let's assume that, based on looking at pilot data, literature and many sources, not what's best for our budget, but actually using information [inaudible], we're going to say that we expect our effect size to be about 0.2 standard deviations. Finally, our sample size constraint. Let's say that the number of students that this program can possibly serve is about 500, but elsewhere in the district or

the state or the area, there's another 500 students that we could assign randomly to a control group.

So, given all of that information, we still have a lot of choice as far as how the study could be designed. So let's talk about the four big areas and the small variations within those four options. First, we can do a simple random sample, right, where we could get a list of all, say, 1,000 students and just sort of ignore what school they go to and all of the rest of it, and we just start with the one list of 1,000 and randomly assign 500 of them to our treatment group. That's more or less a simple random sample design.

We could choose to have no covariants, and we could see what the minimum detectable effect size is, or we could say, you know, these are middle school kids, so maybe there's a seventh-grade or there's a sixth-grade standardized test score that could be available to [inaudible] administrative data, and so what would our sensitivity be if we used that covariant? There's another way we could do it, though, because

these are students in [inaudible] schools, which is analogous to youth in centers or patients in hospitals, but we have a sort of natural cluster that occurs.

So this gives us another three options – complex sample option one, which we'll call between randomization – let's say [inaudible] they pay attention to the school these students go to and assign entire schools to treatment or control, which means we would take a list of the schools we had available and randomly assign half of them to treatment and half of them to control, and whatever school that a student goes to, that student inherits the treatment assignment.

We could also say, let's do this design where we don't have any kind of pretest or other covariant, or, again, we could say, maybe we have a state test available to us, and we want to use that in our analysis. Another way we could do a complex sample is to say, okay, we have our group of schools, and we have students within those schools, but maybe we're

less concerned with this particular intervention of contamination effects. And so it's plausible that we could argue that we can actually randomize students within the school. That means students in the treatment group and students in the control will both attend the same schools, and they might bump into each other and talk [inaudible]. But let's just assume that we're not worried about contamination or anything so that this is a plausible thing to do.

If we do do something like this, we're left with two choices. One is, we could do a so-called random effects model, and in this model the analysis would allow the treatment to differ slightly between all the schools or the sites. This is often an interesting way to go, and this is actually a requirement if we say we wanted the study to generalize to, say, other schools.

Another thing we could do is, we could say, you know what, let's just assume that each school has the same effect, in which case we're going to have a fixed effects analysis. This set of slides really – we're

not going into the multilevel nature of stuff, but, again, in the spirit of that, this is the discussion you want to have with the statisticians, these are your two options.

The fixed effects analysis will generally have more power. It will be more sensitive. And so that seems very attractive from the point of view of what this series of slides are about. However, using a fixed effects analysis does remove your ability to generalize beyond the data in the sample. The fixed effects are basically – you can think of them as sort of a data reduction of, what's the average effect with this group of folks, and you can only discuss it from this particular sample. But in all these scenarios, again, you have situations in which you have covariants or you don't have covariants.

What I'm about to show you now is, we're going to switch to an Excel file, and this Excel file is called PowerUP!, and it's available at the website noted on your screen, CausalEvaluation.org. This resource has been funded by IES, which is the

Institute of Education Sciences. It's well-documented. There's papers published on it. It's a really good source for doing power for these randomized trials. And it will also provide insights for designs and analyses far beyond the scope of these things – things like subgroup designs, moderator mediation, if you want to unpack whatever the mechanism is – these groups of Excel files really do a lot as far as your power. They're available in Excel, and that's what we're going to be working with. They also have R versions for folks who like to use the R computing environment, and I think they have an online version as well.

We're going to go through these Excel files, and what I'm going to do is, I'm going to basically be showing you how to navigate this Excel sheet, and we're going to be plugging in different numbers to get different numbers. But it's going to be a fairly brisk move through what you saw on the previous slide, which is a ton of different designs. I highly encourage you to download from the web page that you're watching this on the PDF Course Level 3 handout. This will

provide not only the results of what we're going to see in our power analyses, but have a lot of narratives that give more insight into the background of what we're assuming, and it also provides examples of how we would write about these results.

So with that, let's switch to Excel. Okay. PowerUP! is essentially an Excel file, and it's an Excel file in which they've added functionality to it to allow for certain computations that are necessary for power analysis that aren't typically available in Excel. So when you download this and open it, you're going to see a couple of warnings that talk about macros, and these are the built-in functionality for Excel. So consult your own IT departments before you agree to anything, but I can say that I have agreed to it. My computer hasn't blown up. And so here we are.

The general layout of this is, it's a pretty well-organized Excel file, but you'll notice that it has tons and tons and tons of sheets here at the bottom. And you could click on any one of these sheets, and a table shows up in which you have things you can type

in and things that it will tell you. However, there's a lot here. And the names of the sheets are somewhat obtuse and kind of hard to work with. Thankfully, the folks who programmed this have programmed lots of little links, and so we can navigate this entire Excel file much like you would a web page.

Okay. And then, by the way, there's a little citation here so you can understand the fine and smart individuals that put this Excel file together. Okay. So let's start with our simple situation, our simple random sample. And so what we're going to do is, each time we do this, we'll go through the step-through process so you can get a feel for how all this works. So I'm going to click on this.

And we see that it gives us another set of choices, sort of choose a study design. Individual random assignment is the synonym for a simple random sample, so let's click on this. And what we want to do is, our scenario here that we've discussed is, we know what the sample size is going to be, so we want to

see what the minimum detectable effect size is. And as we go through all these examples, we want to keep our eye on [inaudible] what we expect. That's what we think we should be able to see out in the world. So let's click on minimum detectable effect.

And so what we have here is lots of different options. Some of these we can tweak. Some of this, we don't want to tweak. So let's go through some of these assumptions. Alpha level, 0.05. This is what we've been talking about throughout the entire series. Two tails for our tests. Power is 0.8. And so you generally want to leave these light yellow ones alone, because these are the typical conventional assumptions.

Capital P is the proportion of your sample that you want to randomize. Again, typically you're fine to just leave this to be 0.5, which is half, which means, of your entire sample, you want half of them to be treatment and half of them to be control. In some situations, you may find yourself where you can randomize only a third of the population available to

you into treatment, in which case you would change this to one-third or 0.33.

And so let's start here with the total sample size. Remember that when we [inaudible], we have 500 students that can be serviced in the intervention, and we know that we can find another 500 for control. So 500 plus 500 is 1,000, so I'm going to switch this to 1,000. We do not have a covariant yet, so we're going to set this to zero, and when we do talk about the covariant, I'll talk about these. So set these to zero and zero.

And that's pretty much all that we have to type in. it's a relatively simple design. We know we have 1,000 people, and we're splitting them by half into treatment and control. Our MDES, then, is 0.177. So what is the minimum detectable effect size? The minimum detectable effect size is, given everything else we've typed in here, what is the smallest effect size that we have an 80 percent chance of detecting? And in this case, it's, rounding a little bit, about 0.18; 0.18 is smaller than what we expect, 0.2,

which, to me, tells me that this is sufficiently powered. In other words, with this design, if we think that the effect that we're going to get out in the world is 0.2, the fact that our minimum detectable effect is smaller than that – it means that we're a little bit more sensitive compared to what we think. So this is great. I would say this is a good design. I would feel good about this. Let's move forward.

But let's think about this a little bit more. Suppose you do have a covariant, sort of a pretest, and let's say that this one pretest is relatively correlated with the outcome. And let's say it's so correlated, in fact, that if you were to run a mile, the correlation would be 0.7. Well, the R-squared statistic is the square of this correlation, and what it does is, it actually tells you the proportion of the variation in the outcome that's explained by, say, the pretest. And without getting into too many formulas or anything, this is really helpful in power analysis.

So 0.7 squared is 0.49. And we're going to use one math test as the pretest, so we have to add that in, and now we see as a result that our minimum detectable effect size has gone down. We've increased sensitivity, because now we can even pick up an even smaller effect size. So, again, 0.13 rounded is a lot smaller than 0.2, which tells me this design has a pretty good chance to pick up 0.2, and even if in fact the impact of the program was, say, 0.15, you know, lower than we expected, we're still walking in with a pretty good chance of having statistical significance.

So those are the first two scenarios using a simple random sample. So let's click start over. And we go back through our step-through process here. Okay. So let's do a slightly more complicated result. Let's say we have students in schools, and so we're going to do some kind of cluster assignment, and it's a cluster random assignment. So let's start with our click-through process here. And it's a simple cluster random assignment. So we'll click on this.

And we have all sorts of things. This PowerUP! is given for more than just two levels. You know, three levels could say – let's say you have schools, but you want to randomize classrooms within the school, and so you want to assign entire schools, but you want to control for the teacher effect, so four levels, say, could be students within classrooms within schools within districts, right? You know, these things can get complicated fast.

Let's just stick with our two-level example for now, and again, we want to do minimum detectable effects. Again, similar table that we've seen before. Again, the light yellow – we want to leave this alone. And let's start walking through some of these options. So first, we want to talk about the total sample. Now, when you have students in schools, you really have two sample sizes, right? You have the total number of schools, and then you have the typical size of the schools.

Well, in this case, we have 20 schools, and we're going to assume that each school has about 50

students. Multiply these two things together, and you get your 1,000 students that we were talking about before. Again, we want to say half the schools are going to be in treatment and half the schools are going to be in control. That's that capital P. Other parameters when you deal with cluster samples – and again, we're not going into a lot of detail in this – but one of the main things, and in fact one of the parameters that you'll probably read a lot about if you were to google, say, my name, is the interclass correlation.

And this is essentially the proportion of the variation and the outcome that's associated with the cluster level. And so Leslie Kish, who is this classic sampling statistician, when he was writing a lot of the interclass correlation, actually sort of – Greek spelling of rho, R-H-O, and flipped the H and the O, and he has this little quip in one of his books where he calls it the rate of homogeneity. And so that's one way to think about what the ICC is.

So what we're saying here is, we think that 15 percent or a proportion of 0.15 of the middle school math scores set as a dependent variable can be associated with the school that they go to. And so this has implications, but [inaudible] about where those implications come from. So we're going to set that there. For now, we're going to assume that we don't have any covariants, so we're going to zero all those out. And that's our result.

So, given the fact that it's the same 1,000 kids, but now we're randomizing entire schools to treatment or control, our minimum detect size has actually shot up quite a bit. It was 0.18, right? And now it's 0.54, which is far larger than what we expect, of 0.2. What this tells me is, this design is on shaky ground, because, because we're randomizing entire groups of students into treatment or control, and not every individual has a shot at it, it changes the statistical model. It's a complex sample. The students are somewhat correlated within the school. That's what that ICC is telling us. And so we actually have far less power than we actually want.

And so if I'm expecting, based on literature, pilot studies, extant information, that my effect size should be 0.2, and I see something that's more than double that for my minimum detectable effect size, I would hit the brakes, gather everybody, gather the budget people, gather the schools and start thinking about possibly a different design. We would need to add, say, more schools in order to achieve our power. And one thing we could do is – if we doubled the schools, we're still greater than it. And if we add 60 schools, still more. And so you can see how these cluster designs can start generating bad news pretty quickly. And so this is why power analysis is a really important thing to do at every stage of the study.

So let's see what happens, though, if we have our pretest. So our pretest now has two statistics associated with it, because if every student has a sixth-grade math score, so some sort of pretest that hopefully is pretty correlated with the outcome, well, what do we do? Not only could we control for

that at the student level, but we could control for it at the school level. Say, we could take the school's average math score from the year before, or we could take the school average of their students' sixth-grade math score.

And so again, let's say that, at the student level, the pretest has correlated with the outcome at 0.7. We square that. We get 0.49. And let's say at the school level – and this tends to be the case, that the R-squareds tend to be higher at schools – let's say that the average pretest of their students is correlated pretty well with the school's average. Let's say that correlation is 0.8. We square that, we get to 0.64.

And we're only really using one variable at the school level, the average, so we'll just put one in there, and now we have a set of 0.5. Our minimum detectable effect size is lower. It's still greater than 0.2, so we probably need to increase the number of schools, but at least we're closer into the

ballpark of 0.2. Still much greater than, so I would call this study underpowered.

So be very, very careful when you start assigning entire schools or entire clusters to treatment and control, because what happens is, you actually lose power pretty quickly. And my little green book and a lot of the other books that I've talked about the end of the slides and I'll talk about at the end of these slides go into some of the mathematical reasons why that's the case.

Okay. For our next example, we're going to say – let's say, having done this cluster level assignment, we've become a little wary of that design because we can't really get more schools or get more students, so what we're going to do is, we're going to figure out a way to make sure contamination isn't a problem, and we're going to randomize the students within the schools. So to do that, let's first say we want to generalize beyond just the schools we have, so we want to do this sort of random effects model. So I'm going to go step-through.

And what we're going to do is, instead of a cluster random assignment, we move over to blocked random assignment. And blocked random assignment comes from the experimental literature. Folks with backgrounds in psychology have probably heard about this type of terminology. But the general idea is that you have blocks, students in blocks, little groups. I think of sandboxes or something. And you basically are doing a bunch of mini-experiments in each of them.

So in this case, the schools would be the blocks, and within each school we're flipping coins or randomizing, and the students within each school are getting treatment or control. So we're going to click on that. Again, we're going to do two levels. And now we have a whole bunch of effects. So we have constant effects, fixed effects and random effects. We'll be going back and forth between the fixed and random effects, but first let's start with random effects.

Again, random effects tell us that our models are going to compensate for the fact that we have some kind of sample of schools and that we want to generalize beyond them. So we're going to, again, click on minimum detectable effect. And a table with light yellow and more intense yellow places we can go through. Okay. So let's fix our blocks and students. So again, we have 20 blocks, 50 students, so for a total of 1,000, and let's go through some of these little parameters here.

The ICC is 0.15, as we said before. We're splitting half and half. We don't have any covariants. Leave that there. And so what we're left with now is this W-looking thing, which is actually the Greek letter omega. This parameter is something that we actually don't know a lot about. There's not a lot of papers that have published the guidance on what omega is. But essentially, what this parameter is is, it's basically saying, okay, if every little sandbox, if every little block has its own treatment effect, that's going to have a certain variation. And then every little sandbox, every little block is going to

have its own average, and that's going to have variation.

And so what this is is, what's the ratio of this variation? We're going to set that to 0.2, which basically says that there's a little bit of a treatment variation. You know, every school's going to have a slightly different difference between the treatment and control. But it's really only 20 percent of how much there's a difference when we think about just the school averages. And so we're going to set that to those parameters, and that's it for the covariants, because we have none in this case.

What we're left with is a minimum detectable effect of 0.207. That's pretty close to the 0.2 that we're expecting. So this is a plausible design. You know, it's technically a little bit underpowered, because in order to get the MDES to be exactly 0.2, I'd probably have to change this [inaudible] 0.79 or maybe 0.78 or – 0.79. Yeah, there we go. So it'd be nice if we could get a couple more schools, but I

think this is pretty close and a relatively [inaudible] situation.

But again, we may have covariants. And so for our covariants, today we're going to use our sixth grade math test, and we're again going to say at the student level, 0.49, and that alone has helped quite a bit. Now our MDES is less than 0.2. But we could also have our covariant at the school level. This R-squared is a little bit different, because it's not really how much of the outcome it explains, but it's how much of the variation in the treatment effect that it explains, right? It could be that schools with lower averages tend to have bigger treatment effects and schools with higher averages tend to have not-as-great treatment effects, and that would create this correlation there. We could say that it's plausible that the correlation coefficient would be such that the square of it would be 0.09, and it has a little bit of an impact, and so what we wind up with is an MDES that's far lower than 0.2, 0.16.

One intuition that I hope you're starting to figure out here is that if you randomize entire schools like we were doing before, you have far less power. If you randomize within schools, and you can keep track of contamination, because the unitive randomization at the individual [inaudible] student, we actually have a lot more power in those situations, even with random effect. So that's an intuition that you can kind of walk away from.

Finally, as our last example, we're going to, again, do a blocked random assignment with two levels, and instead of random effects, we're going to just ignore all this discussion about different sandboxes or different blocks having different treatment effects, and we're just going to say, you know what, we're going to impose onto our analysis – our statisticians are going to force it to all have the same effect, and that's what's generally called a fixed effect.

And so we'll click on this. And our last table of the day – we can start plugging in our parameters. So again, we have 20 schools, about 50 kids a school.

Half of them are treatment, half of them are control. And then our R-squared – we're going to be zero, because we don't have any covariants. No covariants. And there we are. We're getting something pretty similar, actually, to what we got with just the simple random assignment, about 0.18. And so we have fixed effects, and we're acknowledging the fact that each school contains a different set of students, but you don't see any ICCs or anything here. It's all being controlled for.

So, again, this is great. We have a simple-random-sample-like power. But because of the fixed effects, we're constrained about who we can talk about. We can say our study found that, in these schools, we had this effect, and the sentence has to stop there. It can't say, if this were to happen in other schools, we could possibly see these effects. You can't say that with a fixed effects analysis.

And then finally, for our final example, what we're going to do is, we'll add our covariants in. Again, 49ers, and one covariant, and our minimum detectable

effect size has gone down to about 0.13, as before. So I know this has been a complete whirlwind, and so what I encourage everyone to do is to take these last few minutes that I've went through all these tables not as a step-by-step per se, but just a way to get oriented to how tools like this work. And I encourage you to download the handout and read more about what these parameters mean. There's also a lot of really nice sources out there that I mentioned, the books in the back, and then a lot of other agencies give guidance on these things.

So like I said, we've done this whirlwind set of examples, and I've made a table here or wrote down the results of all the examples, and the table is about a four-by-two, so there's eight examples, and again, I just care less about the specific numbers here, because these are totally just made-up examples, but I just want you to look at this table, and the main takeaway is, the choices you make in design, simple versus complex, and the choices you make in analysis, fixed versus random effects, really have an impact on power. And so if you think about

those diagrams and those triangles that I was showing about how study design interacts with analysis interacts with the impact to produce power, this really gives that sense about all those different possible situations.

So now that we've done all this work and we've thought this through and played with budgets, and our lovely set of colleagues that we dearly love but we're kind of sick of because we're still in this proposal phase – now that we've done all this work, now you've got to write it up. And so I've often found as a reviewer that I've never really advocated that a grant be given or a research project be funded because I thought the power analysis was completely killer. You know, that's usually not the thing that I find to be a sufficient condition.

But it is a necessary condition. I have found otherwise smart, intuitive, clever proposals do not get funded when the power analysis just doesn't make any sense. And so here I provide – and also in my little green book with more detail – some general

guidance on what I think a power analysis description should include. You'll often find that writing those paragraphs with these elements forces to check yourself on a lot of things.

So first thing is, you should reiterate your study design. And that includes all the information about how many groups, how the groups are formed, the number of data collection points, and just sort of get all that together. And the next thing is, after you have that sentence, you really should have a sentence that explicitly states what statistical procedure you're going to use to analyze the data.

This is important, because when those two sentences in a paragraph don't match, don't line up, don't make sense, then I get a little concerned that the power analysis doesn't make any sense. You know, it's sort of like a cooking show where they're going to say, all right, we're going to bake a cake. Let's start with cutting cheese. I mean, I suppose there's cheese in some cake, like cheesecake, but you kind of get what I'm saying there. And so the statistical

procedure – if it's, say, a cluster random assignment, I want to say something about a mixed bottle [phonetic] or a two-way enova [phonetic] or an HLM or a random effect. I want to see those sorts of phrases used in the statistical procedure, because that's what is mandated by the design that was [inaudible].

Next, these power analyses, all these numbers that we type into these tables, are basically assumptions. They're good guesses about the future or guesses about what we think the population are like. Those guesses should not be guesses. They should be assumptions that come from good information. And so I want to see where you got that information. And so I'm looking for citations to very long and boring papers about interclass correlations written by a relatively nice guy, or I want to see pilot data. You know, I want these assumptions to come from somewhere, and I want to see that for all parameters, whether you have R-squareds about the predictability, other outcomes based on your covariants, and so on and so forth.

On the fourth thing, I want you to write about what power analysis procedure that you used. If you used the PowerUP!, cite it. These folks worked really hard on making a really good tool. They deserve the cite. And I want to be able to see it so that I can download it and check it out. Formulas are also helpful. You know, sometimes you don't have enough room to put in all the Greek, but that also helps. But, you know, just some sort of citation, software used, the formula, and, again, this is a chance to make sure that what you write in this sentence matches Parts 1 and 2, matches your design, matches the statistical procedure, and that the formulas that you used use numbers that you talk about in Part 3, the assumptions.

What's also often nice to see is a sentence or some sort of indication that you thought about the sensitivity. If you had exactly 0.8 power, and it's a cluster random assignment, I want to see what happens if one of your schools drop out. It may not hurt the inference, but it may hurt the power of your

study if a school drops out. So give me a little sense as to how sensitive this power analysis is to real life, say, global pandemics that happen every 100 years.

And then finally – Part 6 – I actually want to hear the results of your power analysis. I'd like to hear your power is 0.8 or your minimum detectable effect size is this or the number of cases that you need to satisfy the other assumptions is such-and-such, and that you've actually presented evidence that you'll be able to collect this much data.

So, again, a paragraph like this, I think, is something that will be a living document while you do your power analysis, and your power analysis is something that you should do while you're planning your study, and so all of these things can go together. You'll probably often find that you'll have a document called "Power Analysis Final_Final_Final I Mean It This Time Version 2." And so just sort of understand that this is going to be revised and checked again and again.

So to summarize everything that we talked about in Level 3, we've talked about a lot of things. We've talked about how power's the chance to find a statistically significant result. We talked a lot about how it's an informed argument, and you sort of anticipate what you think might happen and use other resources to fill in gaps of what you don't know. Power analysis depends on a lot of factors in your study design, sampling, group formation, how you analyze it, control variables. Again, please do this before the study.

Computations for power analysis can be anywhere from straightforward to very, very complex, and they become even more complex when stuff like [inaudible] and waiting and things we haven't really talked a lot about here come in. There are several existing software [inaudible] that are available that can help provide answers and guidance. And so I really encourage you to look for them. And again, like I said before, be wary of websites that talk about sample size estimators that don't ask about things

like effect size, that don't ask about things like interclass correlations. Those are probably resources being used for, say, market research who are just doing polls or surveys.

Again, just some links here. The PowerUP!, which is what we've gone through, is available at CausalEvaluation.org. It has Excel and R versions. It does way more than the stuff that we talked about here, including moderation and mediation analyses. If you have dichotomous outcomes in simple random samples, things like G*Power, I think, are really helpful, plus it has a suite for many other observational studies, not just experimental, regression correlation and so on. And that is a German program, but its interface is in well-spoken English, so for those of you who don't know German, don't worry about it. You can download it and install it, and it's a wonderful piece of software.

Again, resources that we discussed in the other slides – a lot of these books, all of these books I think, are really good. I have them on my shelf. I

made mine bigger because these are my slides. But really dive in and take a look at a lot of these things. They attack the problem the power at different angles and different levels of complexity, and I know I'm the type of learner that, when I'm trying to figure something out, I usually try to find two or three versions of instruction, and that helps me figure out how things work. And then of course there's papers and blogs and all sorts of things out there.

So this concludes Level 3 and actually concludes all three levels. And so I want to thank you for tuning in, watching this.

CARRIE: Thank you so much, Eric, for this informative class on the concept of power and its importance in evaluation, planning and design. I hope everyone attending now understands more about the process for conducting power analyses and how to apply it for informing evaluation design decisions.

As I mentioned at the beginning of this presentation, this class is the third in a series of three classes

on statistical power. So we hope you will take the time to view the other classes, which provide additional context and instruction on the process of estimating statistical power.

END OF FILE