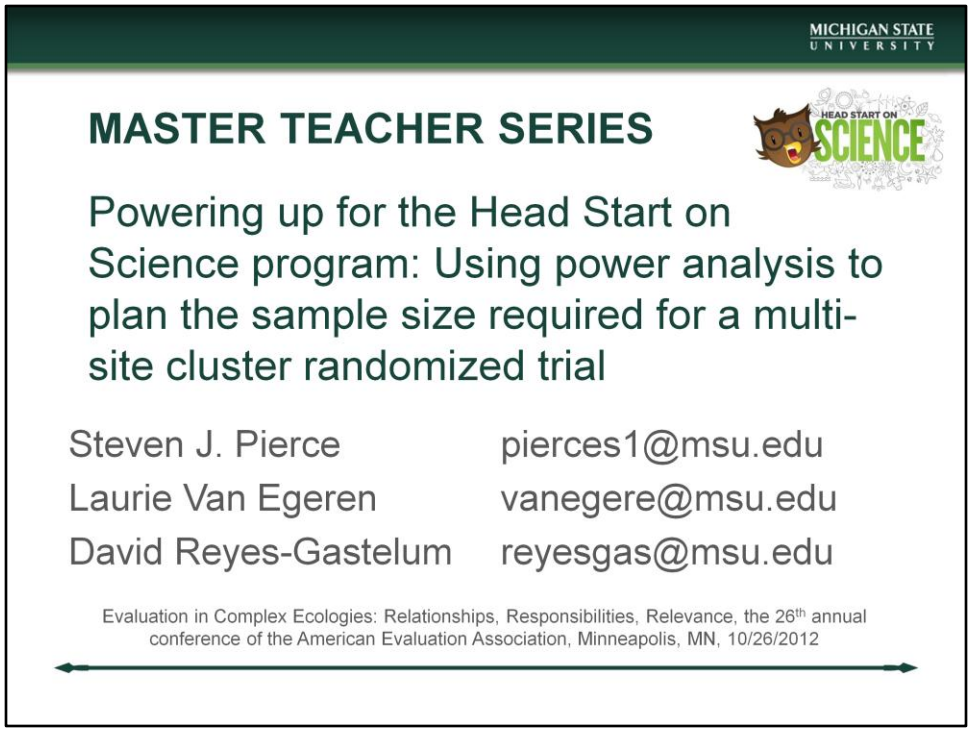


Slides and a resource list will be available from AEA's public eLibrary after the talk.

Demonstration Session #461. Limit ≤ 90 min (75 min talk + 15 min Q&A)
http://eval.org/search12/session.asp?sessionid=7073&presenterid=2097

Citation

Pierce, S. J., Van Egeren, L., & Reyes-Gastelum, D. (2012, October). *Master Teacher Series—Powering up for the Head Start on Science program: Using power analysis to plan the sample size required for a multi-site cluster randomized trial.* Demonstration session presented at Evaluation in Complex Ecologies: Relationships, Responsibilities, Relevance, the 26th annual conference of the American Evaluation Association, Minneapolis, MN.

Abstract: This intermediate session will demonstrate how we conducted an a priori power analysis for a longitudinal, multisite cluster randomized trial of an early childhood science education program, then later revised it to accommodate budget changes suggested by the funder without compromising the viability of the study. We will cover how the research questions, design, and analysis plan informed the power analysis approach; the software we used; and what the input parameters required actually represent. Then we will discuss how we used both pilot data and relevant literature to choose sensible values for those inputs; the potential impact of varying those inputs; the assumptions we had to make; and how we accounted for likely consent rates and levels of attrition. We will emphasize throughout how the audience members may apply the process to planning their own large-scale evaluation studies to increase the ability to document results and improve competitiveness of proposals.


HEAD START ON
SCIENCE
Cluster randomized trial of the efficacy of early childhood science education for low-income children
Laurie Van Egeren
Norman Lownds
Christina Schwarz
Hope Gerde
Holly Brophy-Herb
Steven J. Pierce
Bradley Morris
National Science Foundation
Award # DRL-1119327
NSF

Speaker: SJP

Outline

Head Start on Science (HSOS) study
Power analysis overview
Initial power analysis for grant application
Proposal feedback from NSF
Revised power analysis to win the award

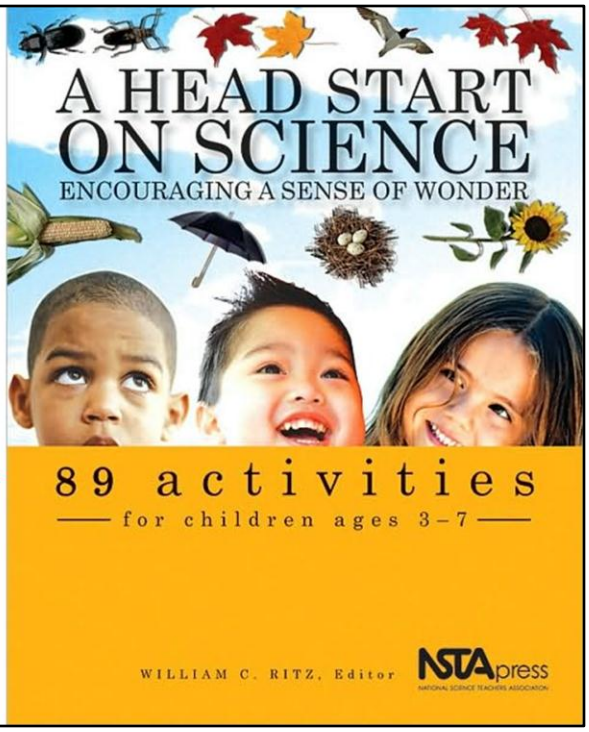
Speaker: SJP

HEAD START ON SCIENCE

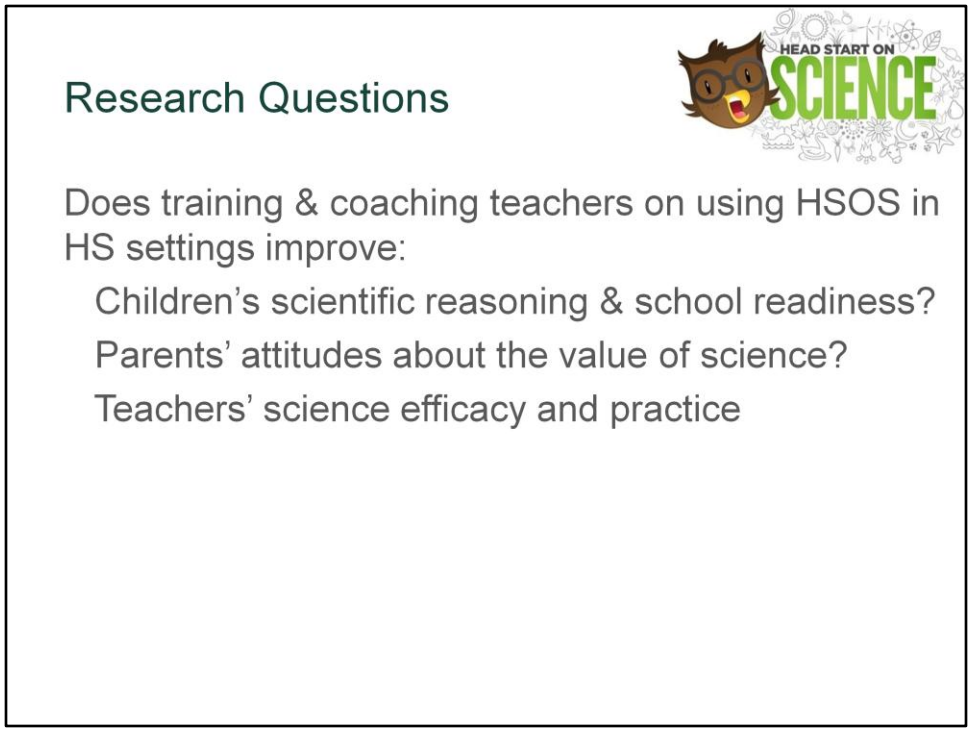
Aligned with:

- National goals
- HS performance indicators

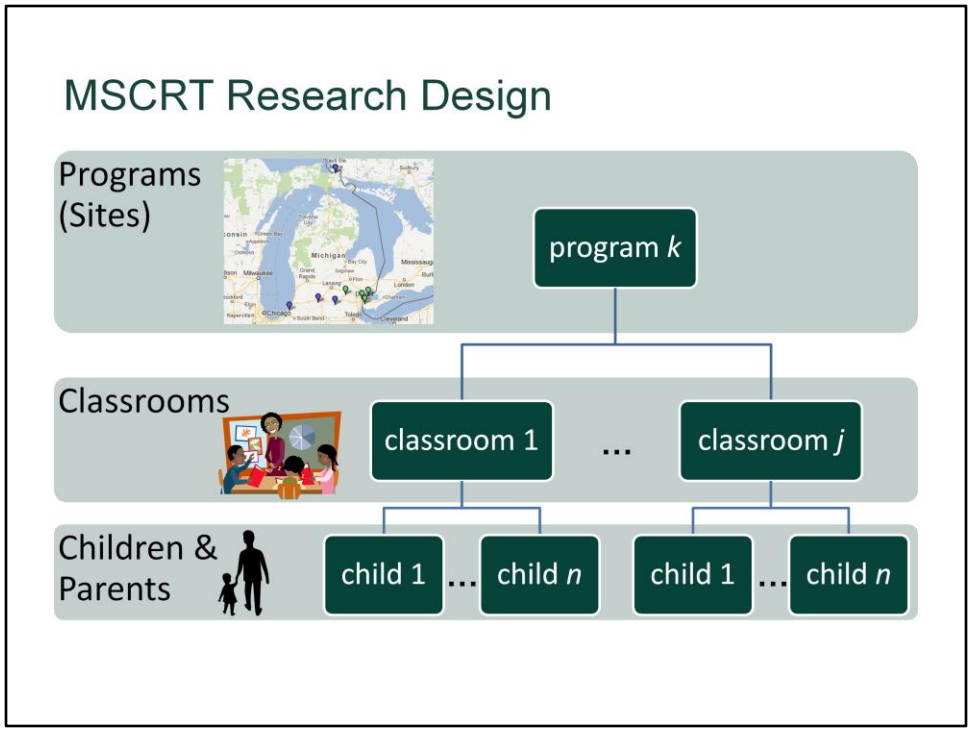
Widely disseminated, but inadequately evaluated!



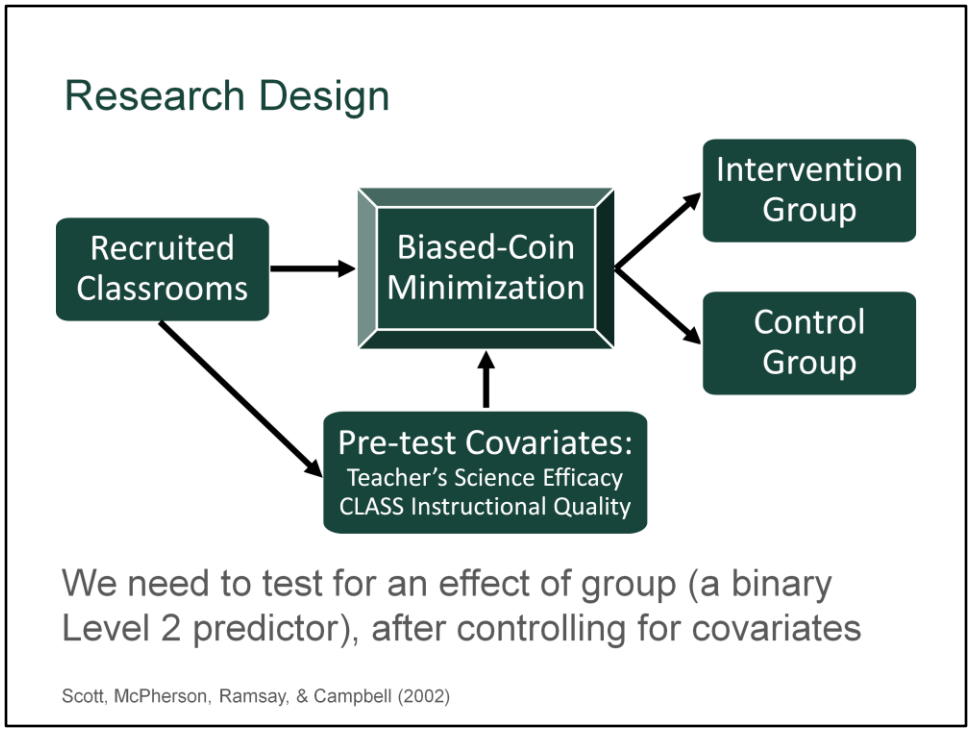
Speaker: LVE
Brief comment on the fact that this is an early childhood science education curriculum that is aligned with national goals and Head Start performance indicators, that it is widely disseminated but has not been adequately evaluated.


Research Questions
HEAD START ON SCIENCE
Does training & coaching teachers on using HSOS in HS settings improve:
Children's scientific reasoning & school readiness?
Parents' attitudes about the value of science?
Teachers' science efficacy and practice

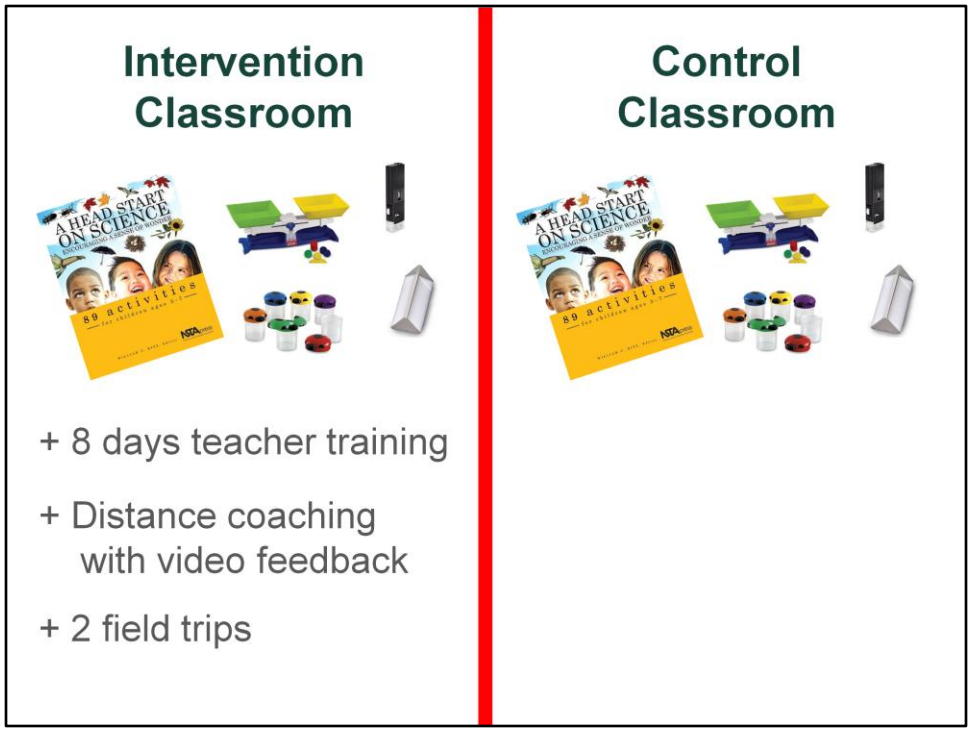
Speaker: LVE



Speaker: LVE?
Because our intervention is provided to teachers, it is inherently a classroom-level intervention. To study the effects of HSOS implementation, we therefore need a sample containing multiple classrooms. However, no one Head Start program was likely to have enough classrooms available to run the whole study with only one program. Furthermore, drawing from multiple programs increases the diversity of settings represented in the study and should therefore increase generalizability of the results. So, we planned to draw classrooms from multiple Head Start programs around Michigan. Finally, the goal is to evaluate the intervention's effects on pre-school children, so we planned to sample multiple children from each classroom. Together, this means the study can be described as a multi-site cluster randomized trial in which programs serve as the sites at Level 3, classrooms are the clusters at Level 2, and children are the level 1 sampling units.



Speaker: LVE
We used a biased-coin minimization process as an alternative to pure random assignment because it does a better job of balancing known covariates across groups when sample sizes are small or moderate. However, this means that the covariates used in minimization must also be integrated into the analyses to preserve accurate inferences about the intervention effect of group.



Speaker: LVE
Both groups received HSOS curriculum materials, so the difference between them is that the intervention group also received intensive training, plus distance coaching via feedback on video clips of their teaching, plus 2 fields trips. Thus, the group effect represents whether or not providing intensive professional development to the intervention teachers is beneficial.

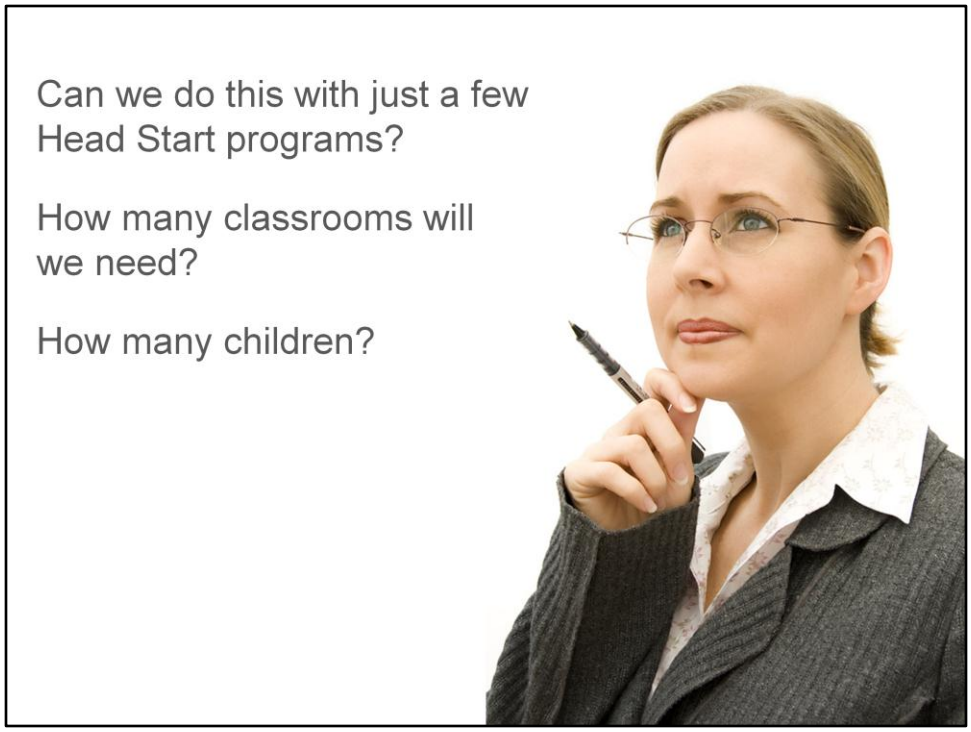
Research Design

Group	Longitudinal Design			
Intervention	O_1	R	X	O_2
Control	O_1	R		O_2
	Pre-test	Assign Group	Intervene	Post-test

Using longitudinal data should strengthen causal inferences about the group effect

Shadish, Cook, & Campbell (2002)

Speaker: LVE
Using longitudinal data in a randomized trial like this strengthens the design and improves our ability to make causal inferences about the effect of the intervention. We're saying this here because it sets the audience up for technical parts of how we plan to analyze the data later, which affect how we did the power analysis.

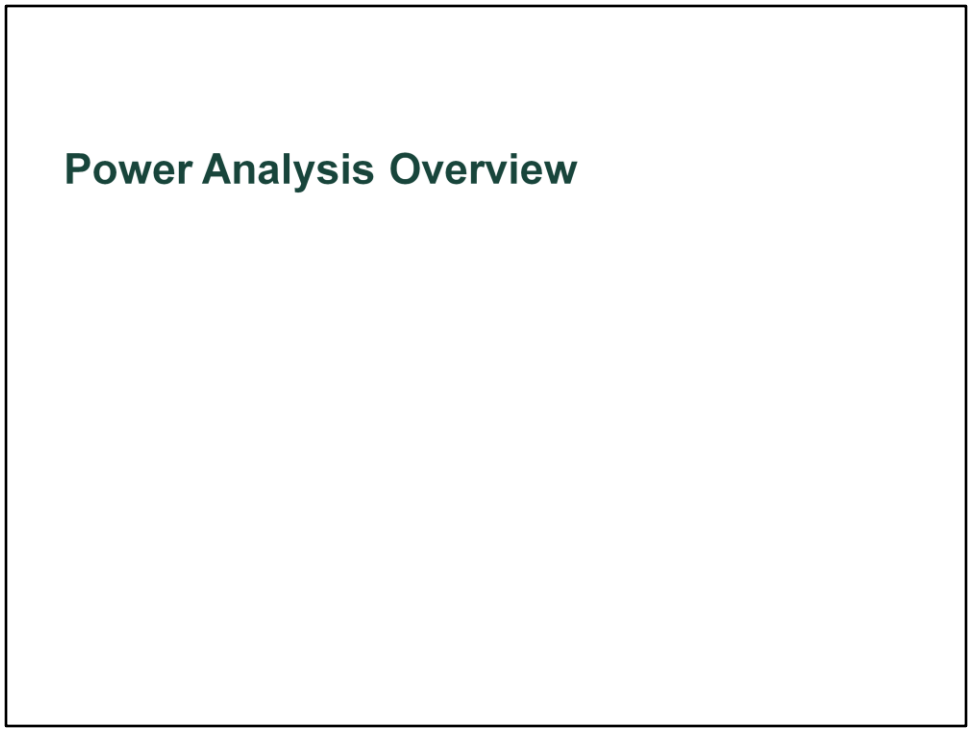


Speaker: LVE

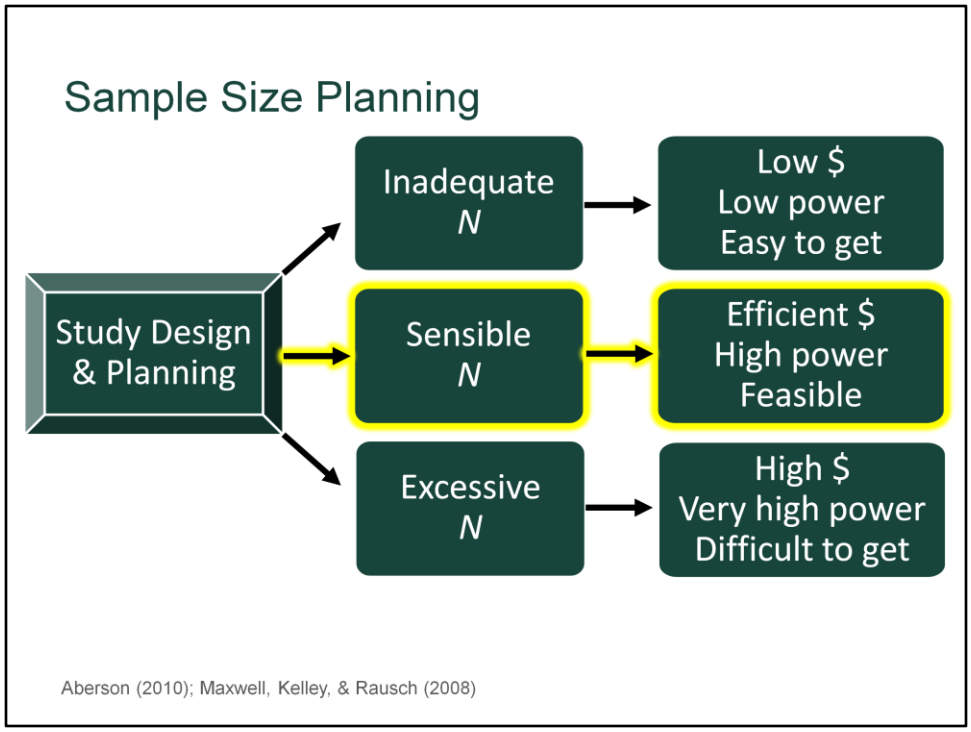
These are study design questions that a statistician can help answer by doing a power analysis.

Speaker: LVE

LVE had an initial idea about sample sizes (4 programs x 10 classrooms each x 6 kids each = 40 classrooms & 240 kids), but wanted to confirm that would be sufficient. That's something we can check with power analysis.
After this slide, we transition to SJP as the speaker.



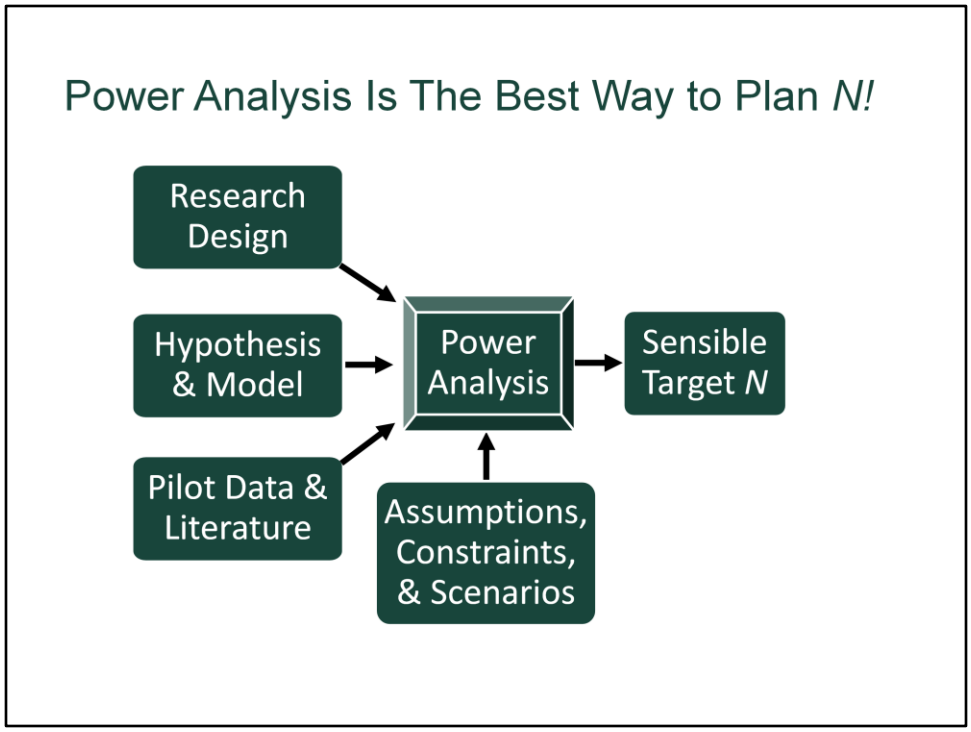
Speaker: SJP
So Laurie just gave you the context surrounding our Head Start on Science project, now I'm going to start connecting that context to the technical material on power analysis and sample size planning. First, I'm going to provide a brief overview of power analysis to set the stage, then I'll go on to discuss the nuts and bolts of how we conducted the power analysis for this study.



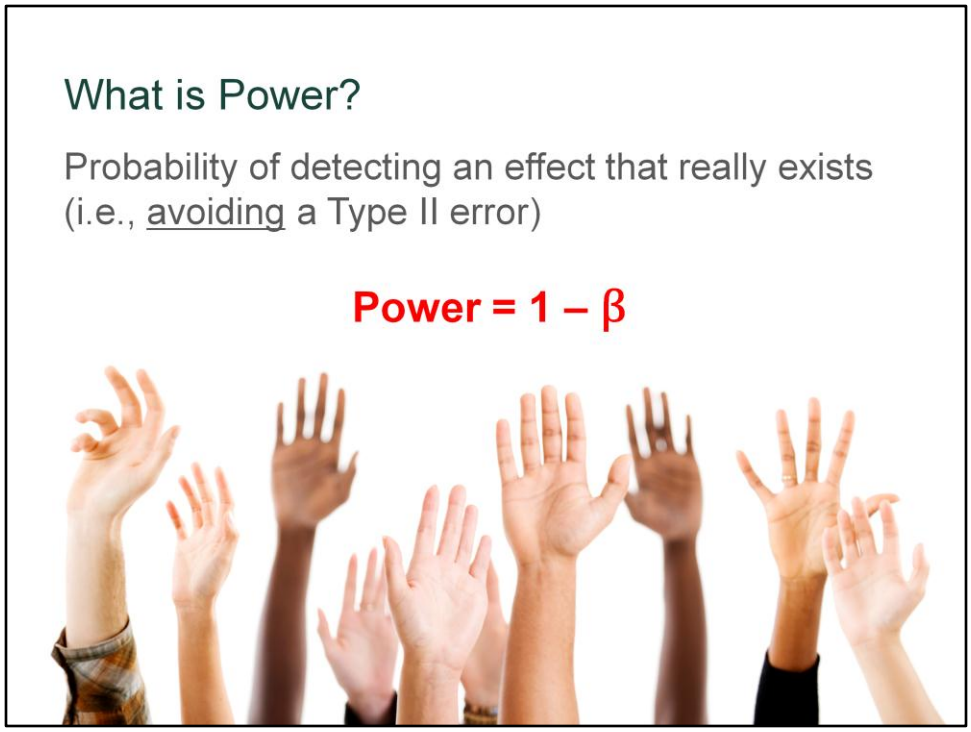
Speaker: SJP
There are 3 possible outcomes of sample size planning at the research design stage for a study, 2 of which have undesirable consequences. First, you could end up making decisions that lead to inadequate sample size (one that is too small). While it is cheap and easy to recruit a small sample, that will give you low power to detect real effects of interest.

Second, you could make decisions that lead to an excessive sample size (N too large). While that provides very high power to detect real effects, the downside is that it will be very expensive and difficult to obtain.

So the sample size planning goal is to find that third outcome: a sensible and appropriate sample size that provides high power to detect effects, while remaining efficient with respect to cost and feasible to collect. How do we get there?



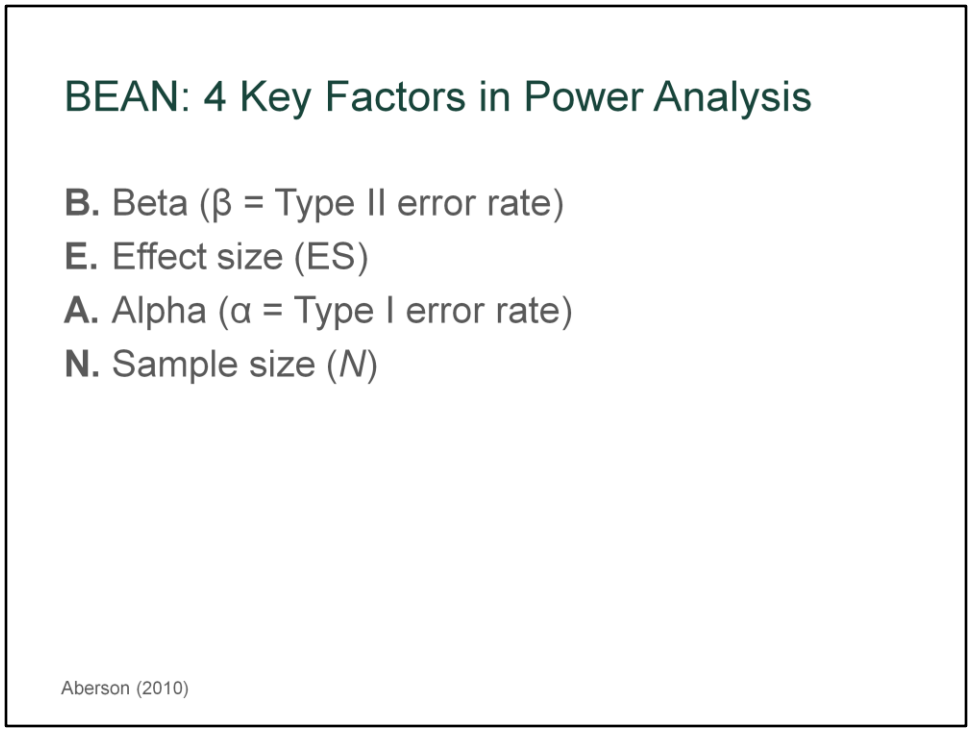
Simply put, power analysis is the best way to plan the sample size for a proposed study because it will tell you what target sample size is sensible. But, to do it well, you need to carefully consider a number of inputs. Perhaps the most important inputs are the research design, the hypothesis you want to test, and the statistical model you will use to do that test. They're crucial because you need to pick a power analysis method aligned with the statistical analysis you plan to run. In addition, you need to use pilot data and/or prior literature to inform your work, plus you may need to make a few other assumptions and consider possible constraints and scenarios. That's what I'm going to illustrate for you today with a concrete example based on a real project.



Before we go much further, it's useful to stop and ask "what is power?" Can anyone give me a 1 sentence definition of statistical power?

Power is just the probability of detecting an effect that really exists, which means that it is also the probability of avoiding a Type II error. We use the symbol beta for the probability of making a Type II error.

Power has a very simple relationship with beta. Assuming the effect is really there, you can either accurately detect it or you can make an error by failing to detect it. The sum of the probabilities for those two possibilities must equal 1, so power equals one minus the Type II error rate. If you know beta, then you also know power. You want power to be high because you want to be pretty sure that you will detect an effect when it is really there.



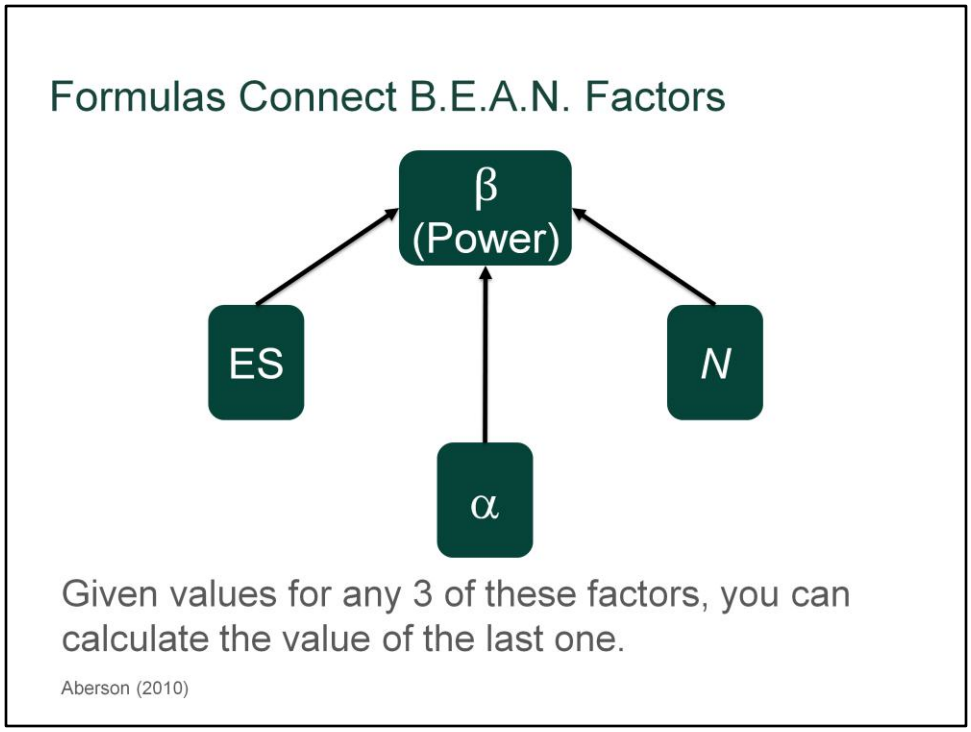
The acronym BEAN nicely captures four key factors that are part of every power analysis in one way or another.

The B in BEAN stands for β, which is the Type II error rate. I just mentioned that subtracting β from 1 gives you the expected power, so if you know β, then you know how much power your study will have. Reducing β increases power.

The E stands for effect size. If all else is equal, you will always have more power to detect large effects than you do to detect small effects. The appropriate measure of effect size depends on the specific kind of power analysis you need to do.

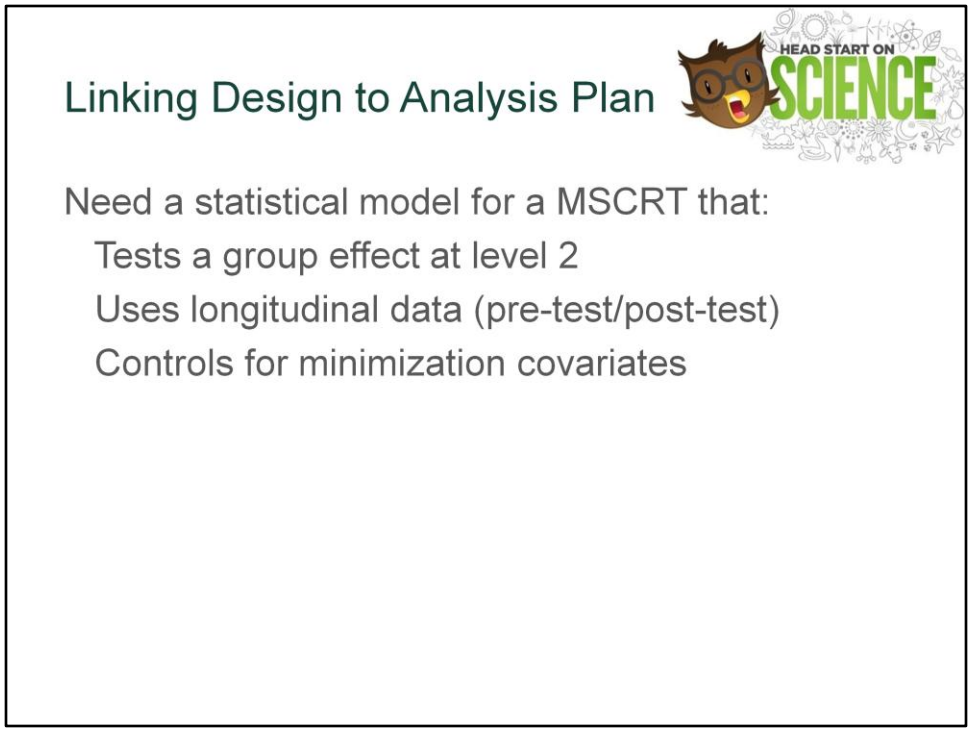
The A stands for the significance criterion α, which is the Type I error rate you are willing to risk. Increasing α lowers the bar for how much evidence you want before you decide there is a significant effect. That gives you more power, but increases the risk of falsely concluding there is an effect when there really isn't one.

Finally, N stands for the sample size you use in the study. If all else is equal, larger samples give you more power to detect real effects than do small samples.

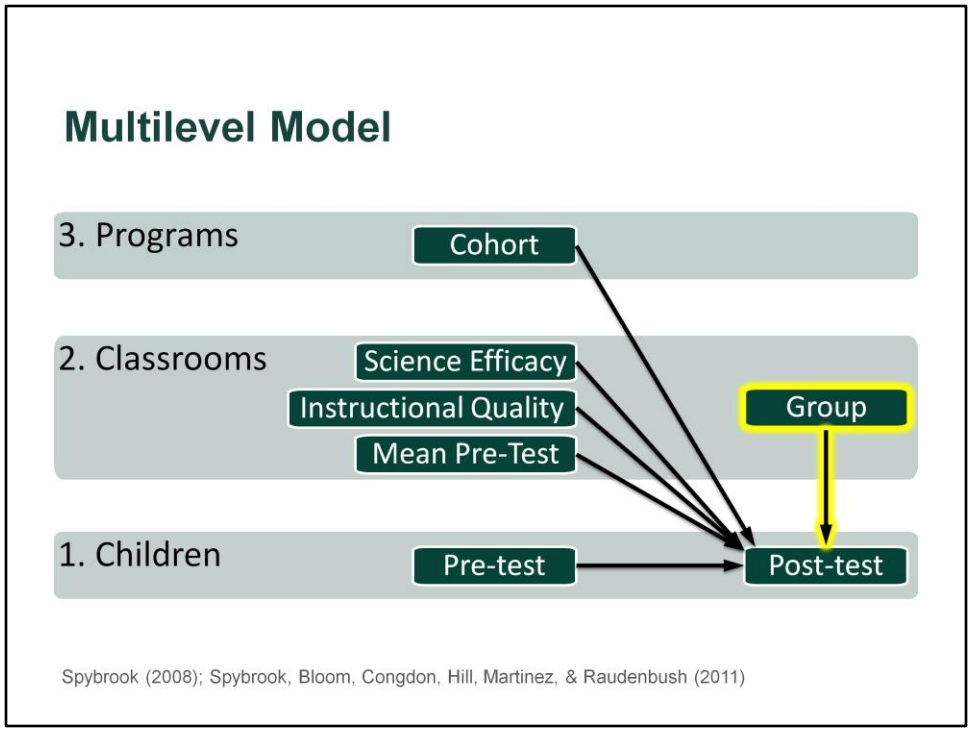


There are formulas associated with each statistical model that show how these four factors (and possibly a few additional factors unique to the statistical model) are mathematically related to each other. So, given values for any 3 of these factors, you can use those formulas to calculate the value for the remaining one. Power analysis is the art of using those formulas to understand and exploit those relationships between factors to plan an efficient, adequately powered study.

I'll be demonstrating one of the most common ways to use power analysis, which is to calculate how much power you would have given a particular effect size, the alpha level, and a particular sample size. You can think of those assumptions as the legs of a tripod that support a conclusion at the top: namely the power that the study would have to detect an effect of the specified size if it really exists.



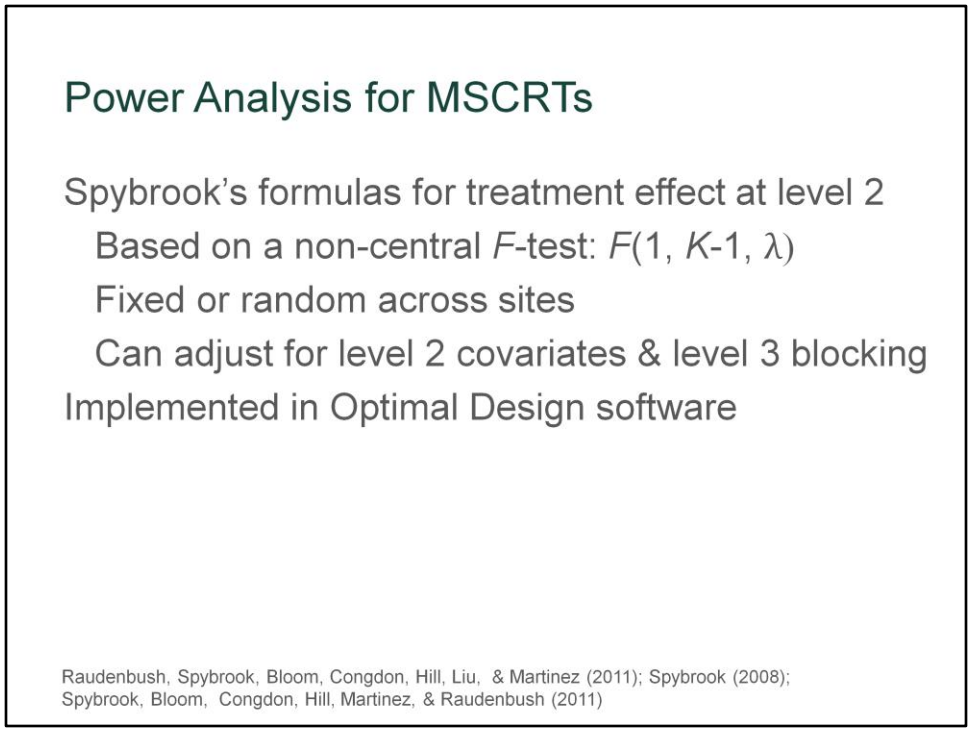
Speaker: SJP
Recall that this is a multisite cluster-randomized trial, with children at level 1, classrooms at level 2, and programs at level 3. That puts some boundaries on the analysis we might need to perform to answer our research questions, but we still need to flesh out the analysis plan before we can identify a suitable power analysis approach. We need an analysis for a MSCRT that will actually test the group effect at level 2, effectively uses the longitudinal pre-test/post-test data we plan to collect, and that also controls for the minimization covariates that we need to incorporate into the analyses because of how we did the group assignment.



Speaker: SJP
This diagram illustrates the multilevel model at the heart of our analysis plan. We will test for a group effect on children's post-test scores, after controlling for children's individual pre-test scores, classroom-level mean pre-test scores, the two minimization covariates (Lead teacher's science efficacy and CLASS Instructional Quality), and a program-level cohort effect. We assume that the outcome will be a continuous variable.

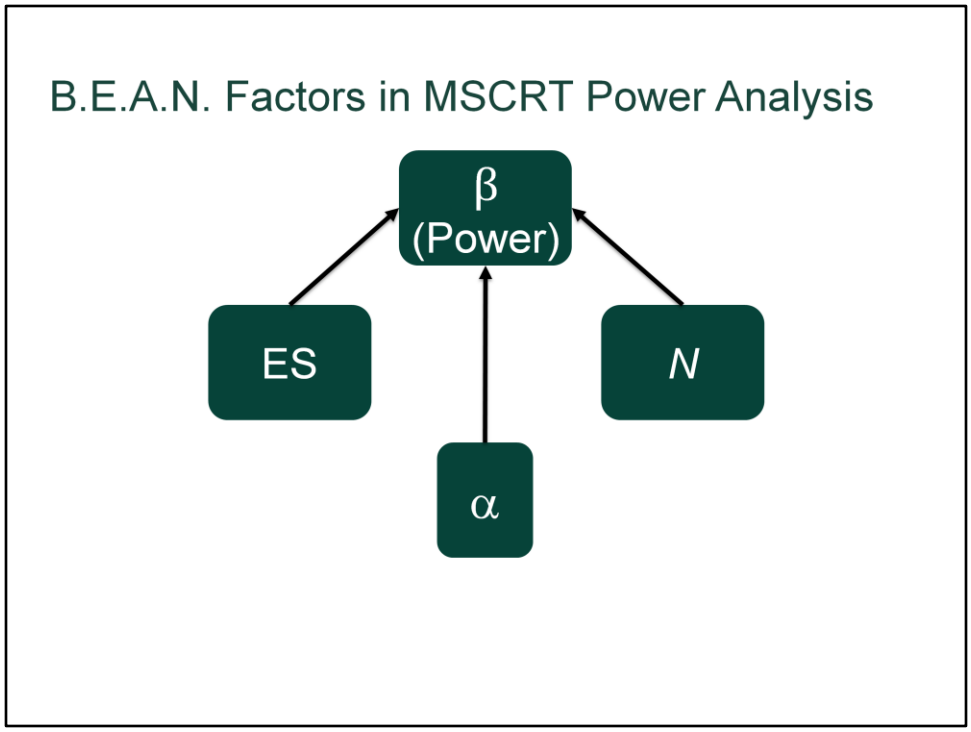
By incorporating all three levels of analysis and placing the group effect at level 2, we align the analysis with the MSCRT research design. Incorporating the individual pre-test as a predictor takes advantage of the longitudinal data, while adding Science Efficacy and Instructional Quality controls for the minimization covariates used in group assignment. I'll explain why we added the classroom mean pre-test as a level 2 variable later.

Now that we've identified a statistical model for testing our hypotheses, we need a power analysis methodology so that we can figure out how to measure effect size.

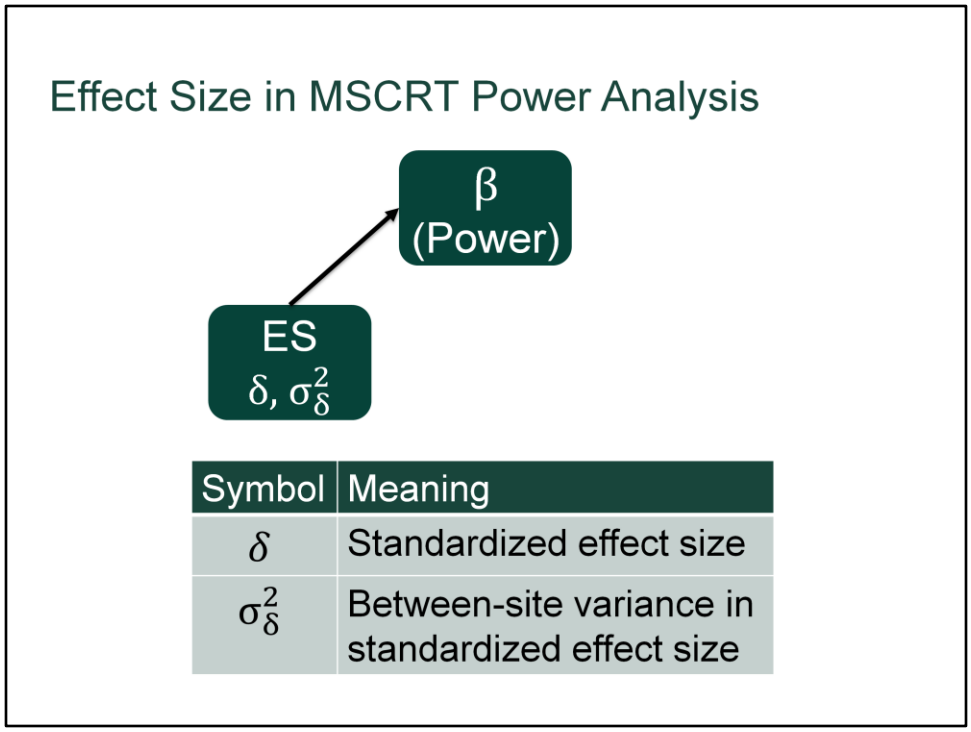


I searched the methods literature and found that Jessaca Spybrook's work on power analysis for multisite cluster-randomized trials matched up with what I needed to do very well. She's building on work by Raudenbush, Bloom, Hedges, and a few other researchers who write extensively about multilevel models and cluster-randomized trials. These formulas focus on power for a binary treatment effect at level 2 using a non-central F-Test. The treatment effect can either be fixed or random across sites, and the formulas allow us to adjust for level 2 covariates and blocking at level 3.

It is particularly helpful that these formulas are implemented in a free piece of software called Optimal Design published by Raudenbush, Spybrook, and colleagues. I'd also like to note that Jessaca is actually also presenting on power analysis for cluster randomized trials in Session 619 at 2:40 PM today. Her session was also selected as one of the Master Teacher Series sponsored by the Quant TIG. She may go into more technical detail or cover a broader perspective on cluster-randomized trials than I aim to provide because I'm demonstrating how we applied these tools to a particular project.



So, now that we have identified a power analysis method appropriate to the planned analysis, we need to understand how it works. Let's start by identifying the input parameters in Spybrook's formulas that correspond to the BEAN factors I mentioned earlier. We're going to need values for the effect size, the alpha level, and the sample size at the very least, plus there are a couple other input parameters as well.



In this method, there are actually two different parameters for the effect size, delta, which represents the standardized effect size, and the between-site variance in the standardized effect size. The latter parameter allows you to assume that the treatment effect may actually vary randomly across sites, which could be substantively important. If that variance is set to zero, you're assuming the treatment effect is constant across sites. Let's now look at how delta is defined.

Effect Size in MSCRT Power Analysis

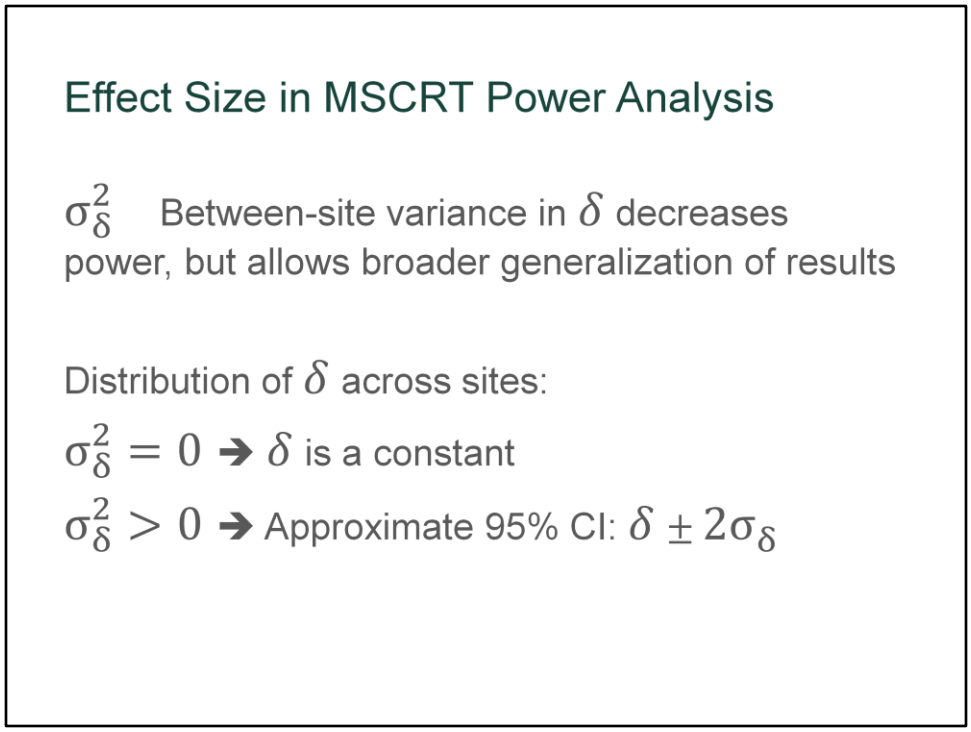
$\gamma = \overline{Y_T} - \overline{Y_C}$ group effect (difference in means)

$\delta = \frac{\overline{Y_T} - \overline{Y_C}}{\sqrt{\tau_\pi + \sigma^2}}$ standardized group effect (SD units)

SD based on 2 variance components:

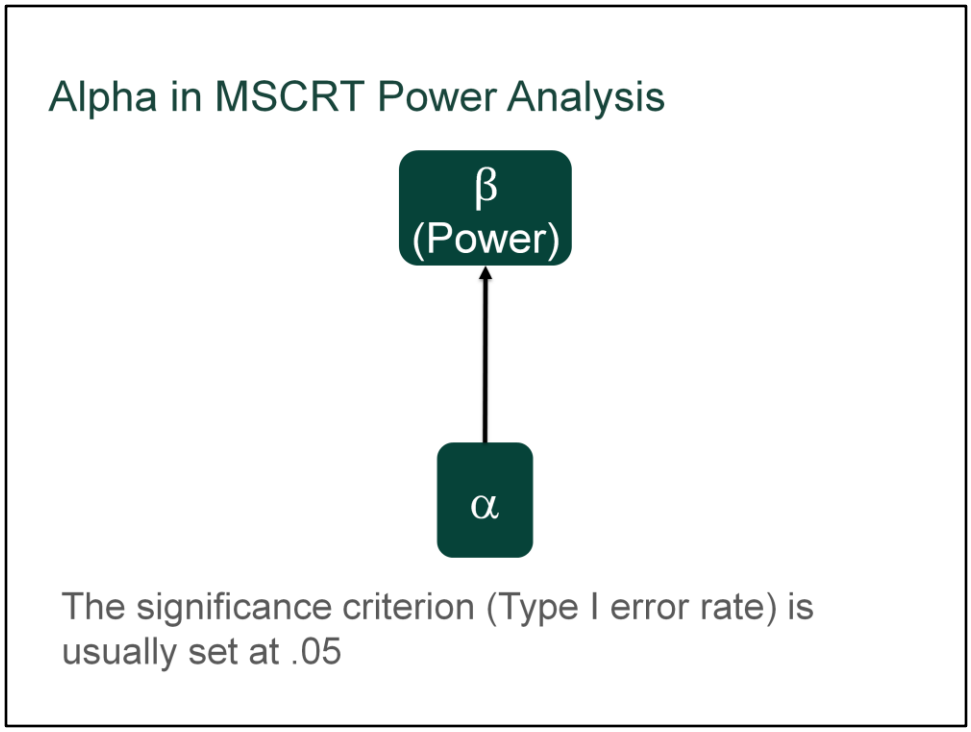
- Between-cluster variance within sites (τ_π)
- Within-cluster variance (σ^2)

In essence, the raw measure of effect size (gamma) is the difference in mean outcomes between the treatment and control groups. However, it is easier to communicate effect sizes in standardized terms, so we calculate delta by dividing that raw difference by a standard deviation based on pooling two variance components: the between-cluster variance within sites, and the within-cluster variance. That means delta is conceptually similar to Cohen's d, which is used as the effect size measure in a t-test. Both are standardized mean differences expressed in units of standard deviations.

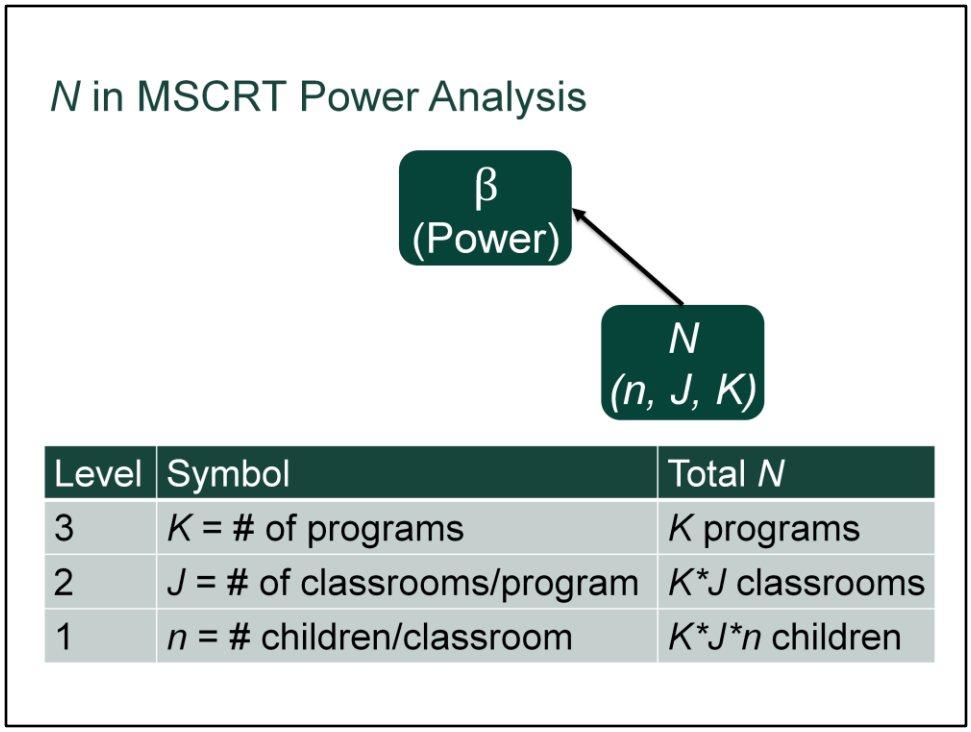


When we assume there will be variability between sites in the effect size, we acknowledge the fact that the intervention may not have the same effect in all Head Start programs due to local contexts and that reduces our power slightly. However, it also broadens our ability to generalize the results to a larger population of Head Start programs rather than restricting our study to inferences only about the effect of our intervention on just the programs we studied.

Setting this variance to zero implies that we think the intervention will be equally effective across sites, so delta will be a constant. Setting it larger than zero implies that we can expect the average effect size be delta and 95% of the sites to have effect sizes within 2 standard deviations of that average.



Alpha, the significance criterion that represents the Type I error rate, is the same in an MSCRT context as it is in other statistical analyses. We usually set this at .05 in the social sciences, indicating a 5% chance of falsely concluding there is an effect when there really is not.



Level	Symbol	Total *N*
3	*K* = # of programs	*K* programs
2	*J* = # of classrooms/program	*K*J* classrooms
1	*n* = # children/classroom	*K*J*n* children

One of the issues that comes up in a multilevel study like ours is that sample size is no longer a single number. You have a separate sample size at each level of analysis. So, I need to introduce some more symbols that I'll be using later on. K refers to the number of programs or sites. J refers to the number of classrooms per program, so that K*J is the total number of classrooms involved. Finally, small n will refer to the number of children per classroom, so that the total number of children will actually be K*J*n. When I want to refer to a whole set of values for n, J, and K, I'll just use the capital N.

N in MSCRT Power Analysis

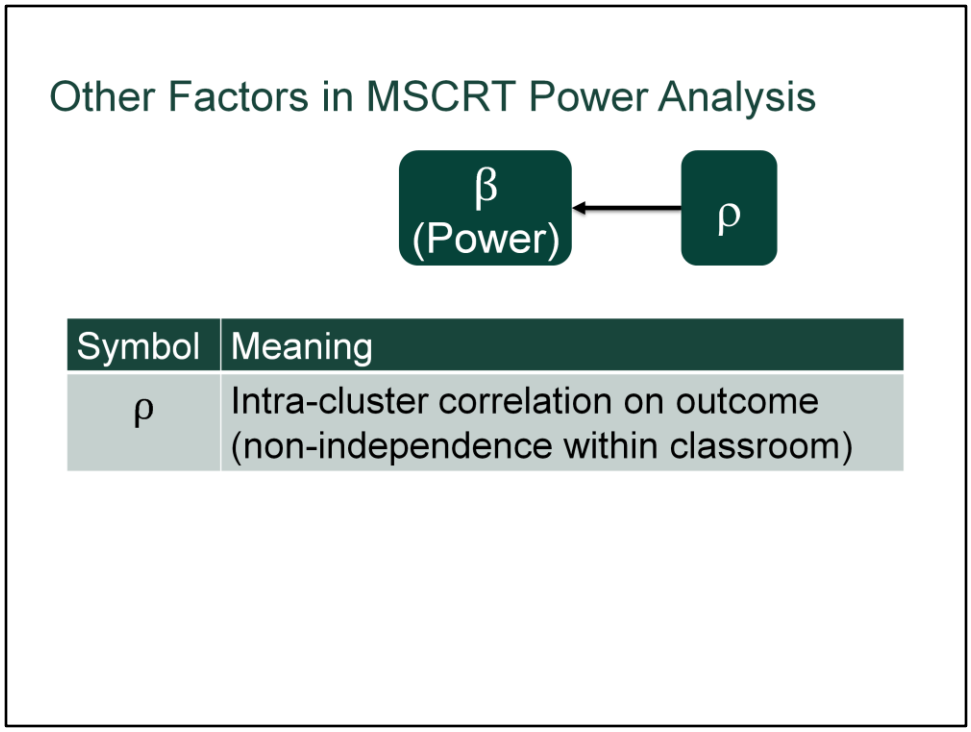
K & *J* affect power more than *n*, but data collection cost per unit may vary across levels

- Adding a program is difficult & expensive
- Adding a classroom expensive
- Adding a child is cheap

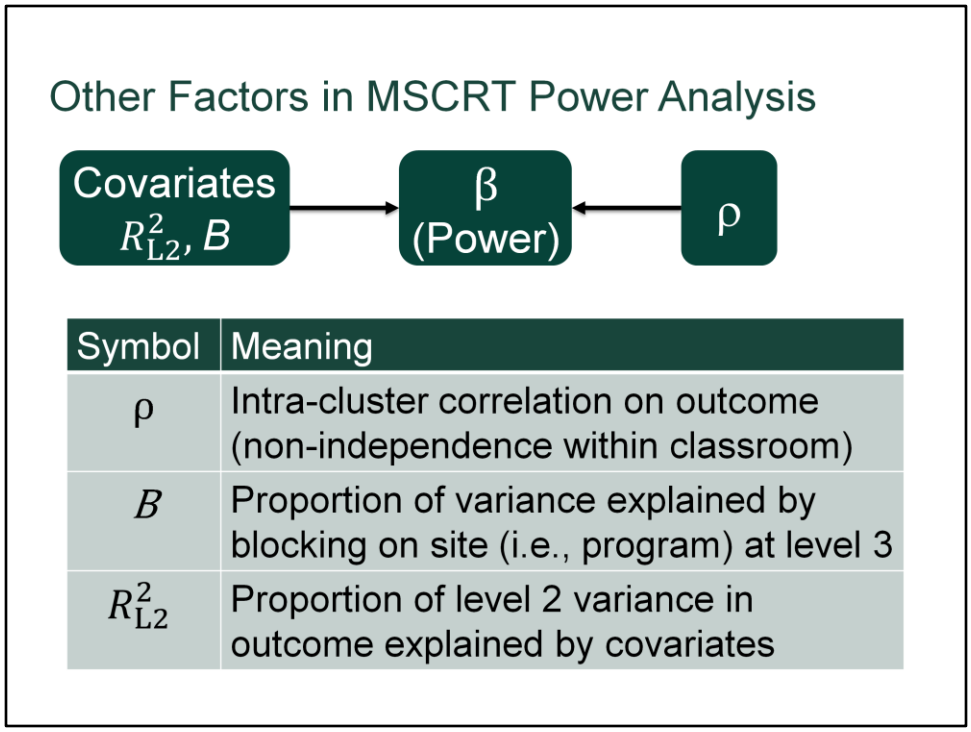
Consider feasibility & cost vs. power for alternative combinations of *n*, *J*, & *K*.

Changing K and J – the sample sizes at the higher levels of analysis – will have stronger impact on power than changing n, but the data collection costs per unit may vary a lot across levels in the design. It may be more costly to add another program than it is to add another classroom or another child. In our case, adding a whole program was both difficult and expensive; adding a classroom was also expensive, but adding a child was fairly cheap.

In multilevel studies, you should consider balancing the feasibility and costs of alternative combinations of n, J, and K against the power they yield.

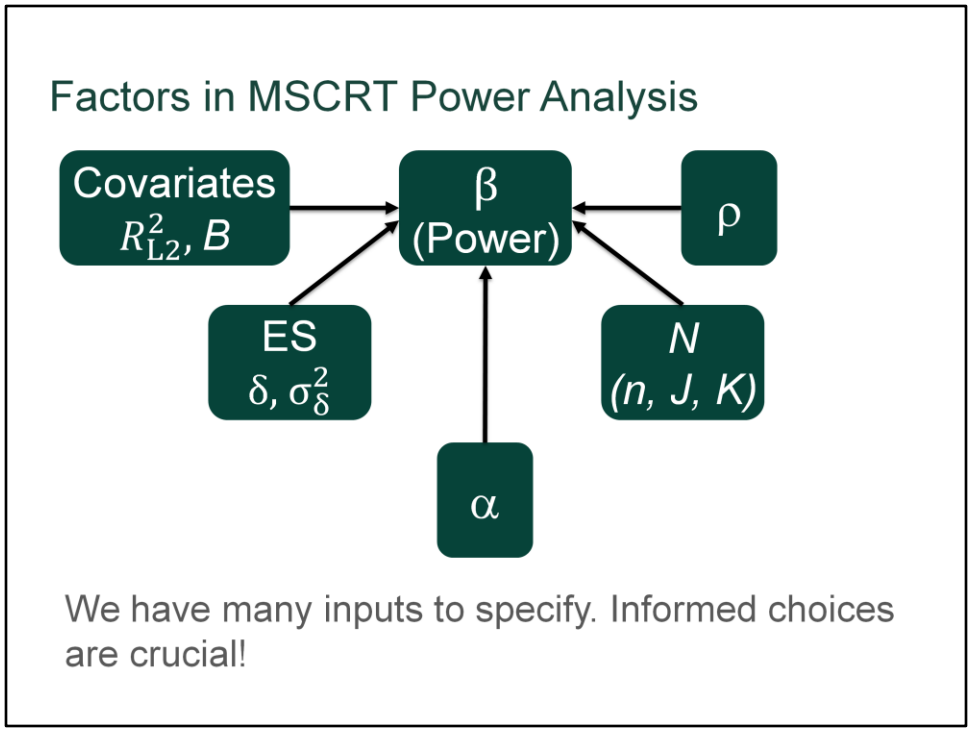


Now, there are a couple other input parameters in Spybrook's formulas that are related to the multilevel nature of the statistical model that we need to consider as well. First, let's consider rho, the intra-cluster correlation or ICC. This represents the non-independence in outcomes among children from the same classroom. One of the fundamental reasons why we run multilevel models for studies like this rather than simple regression models is because regression assumes the data are all independent, but that is rarely true in multilevel studies. Multilevel models correct our statistical tests for this non-independence. Larger values of rho reduce power because you're not getting as much unique new information from each child observed at level 1. Good level 1 covariates (such as the child's pre-test score) can reduce the effective value of rho.



Another factor related to using the 3-level model is that we can account for site-level variance in the outcome as well, which we can do by including blocking covariates. The parameter B represents the proportion of variance explained by level 3 covariates such as site, or site-level characteristics like which cohort a site belonged to. Increasing B will increase power.

Finally, remember that we said earlier there were some covariates at level 2 because of how we did group assignment, plus a classroom-level mean pre-test score. The R-squared here refers to the proportion of level 2 variance in the outcome explained by those covariates. Explaining some of that variance with covariates increases the precision of the estimate for the group effect and therefore increases power.



Collecting all those pieces shows just how many inputs we will need to specify. Making informed choices about what values to use for each one is crucial to doing power analysis well. You should be ready to justify the value used for each input.

F-test Noncentrality Parameter (Lambda)

$$\lambda = \frac{K\delta^2}{\sigma_\delta^2 + 4[\rho + (1 - \rho)/n]/J}$$

Increasing λ increases power
Covariates & blocking increase power through impact on δ, σ_δ^2, & ρ

This is the formula for the noncentrality parameter (lambda) in the F-test for the treatment effect in a MSCRT. I'm not going to dissect it in detail, but I do want to note that as lambda increases, then power increases. You can see several of the input parameters to the power analysis directly in this equation, namely the effect size parameters, the intra-cluster correlation, and the 3 different sample size parameters. Adding covariates and site-level blocking increases power through their impact on the adjusting the effect size and intra-cluster correlation parameters.

Initial Power Analysis

Now that you have an overview of the parameters involved in the power analysis, I want to start demonstrating the actual power analysis process we went through as we prepared our grant proposal for NSF.

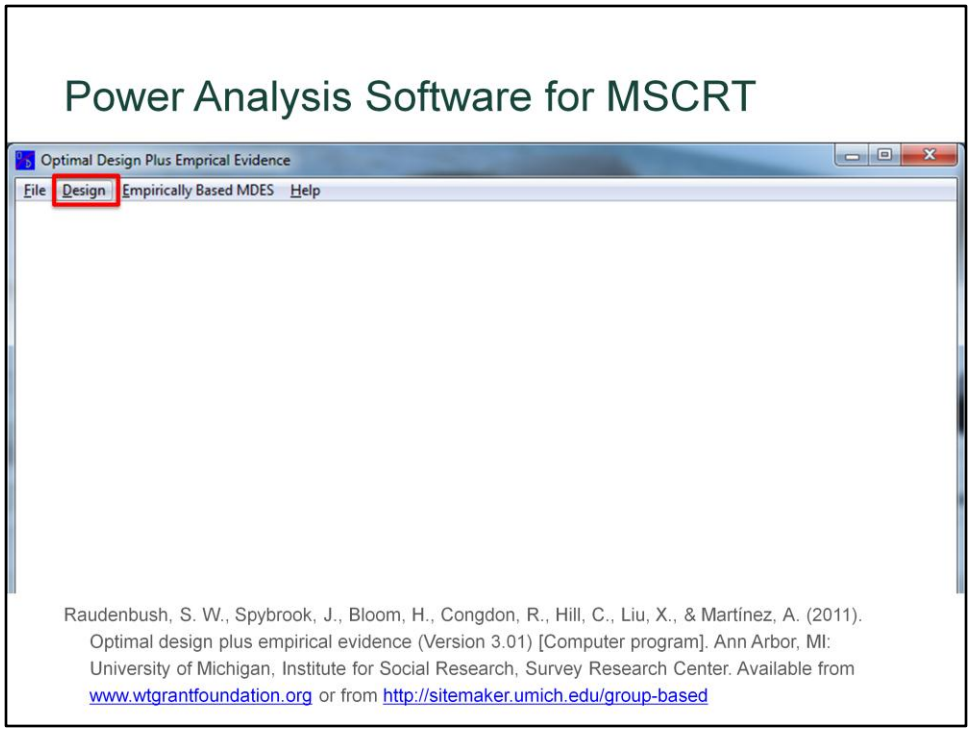
What if we recruit …

K = 4 Head Start programs

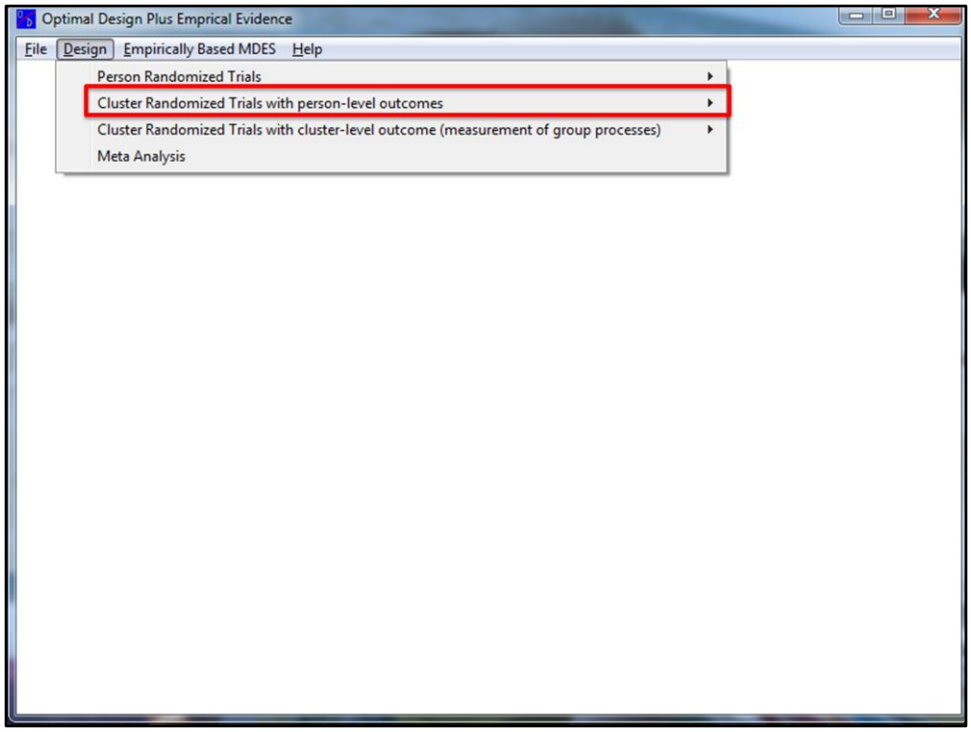
J = 10 classrooms/program

n = 6 children/classroom?

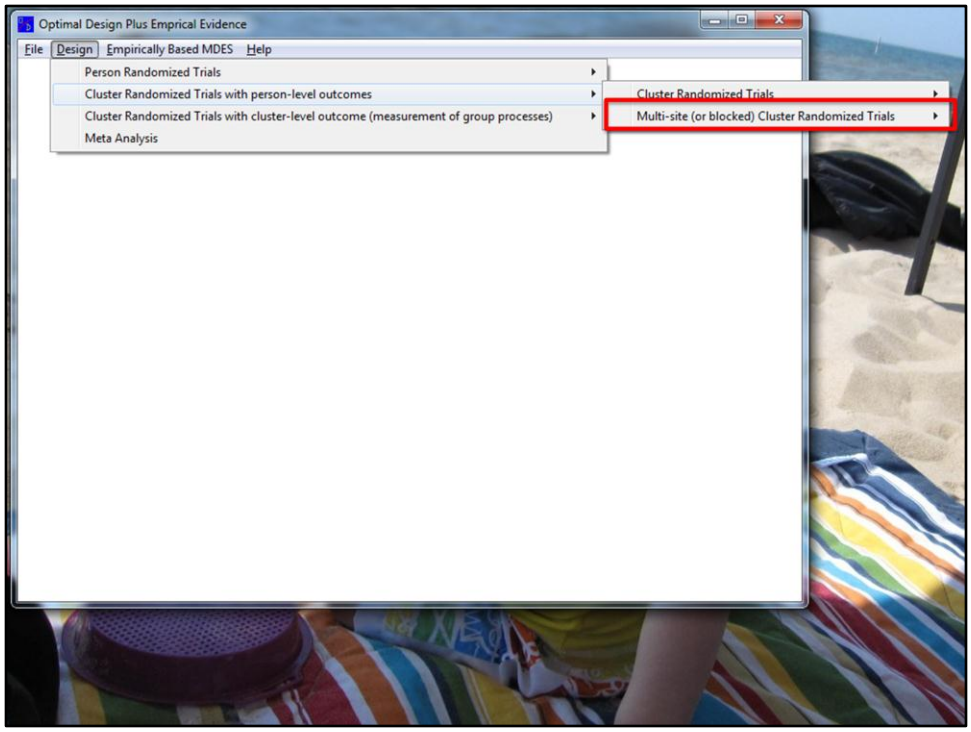
Recall that Laurie had asked me whether we would have enough power if we recruited 4 programs, with 10 classrooms each, and 6 kids per classroom. Let's start plugging numbers into the Optimal Design software to see how much power that would yield.



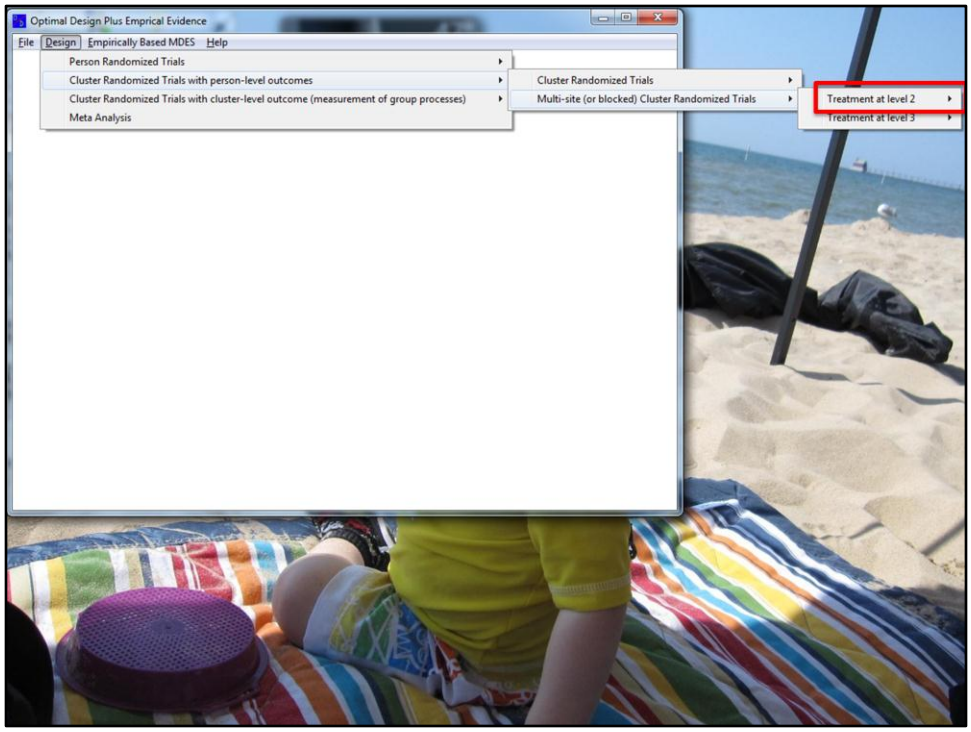
So this is the Optimal Design software, which is a free tool you can download from the web. We using the Design menu to select the research design we are going to use, so click “Design”.



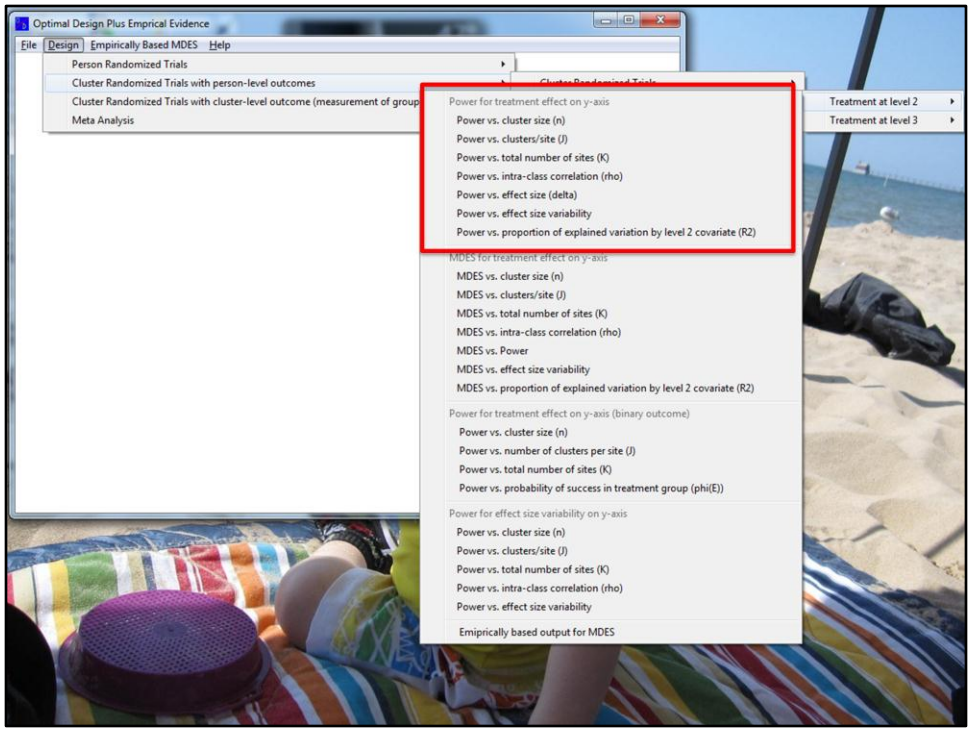
Then click “Cluster Randomized Trials with person-level outcomes.



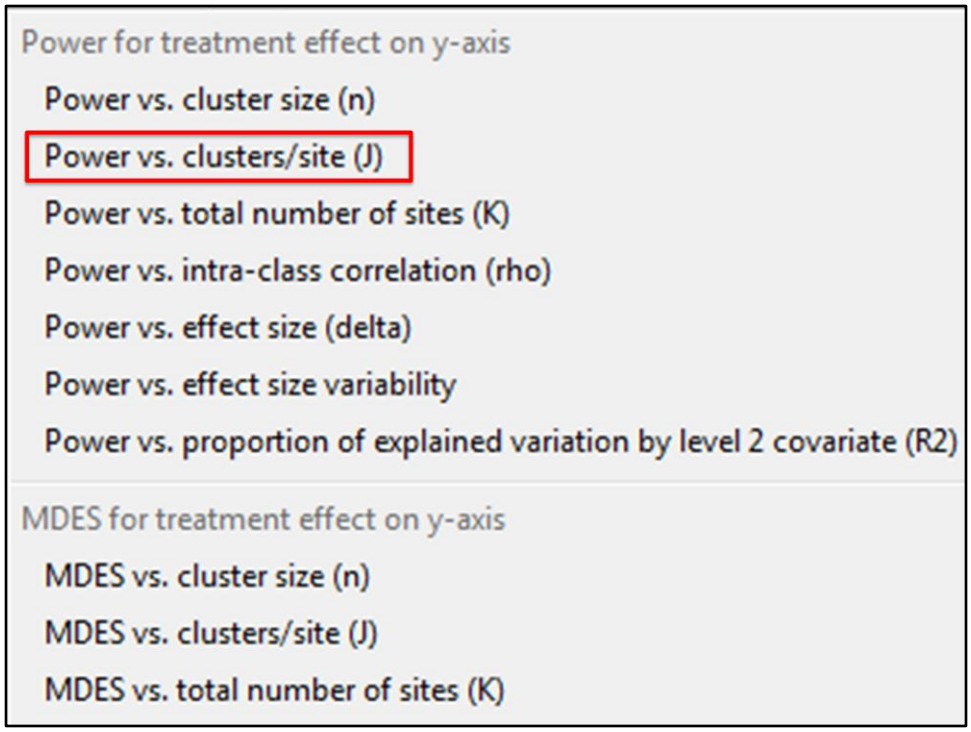
Then click “Multi-site (or blocked) Cluster Randomized Trials”



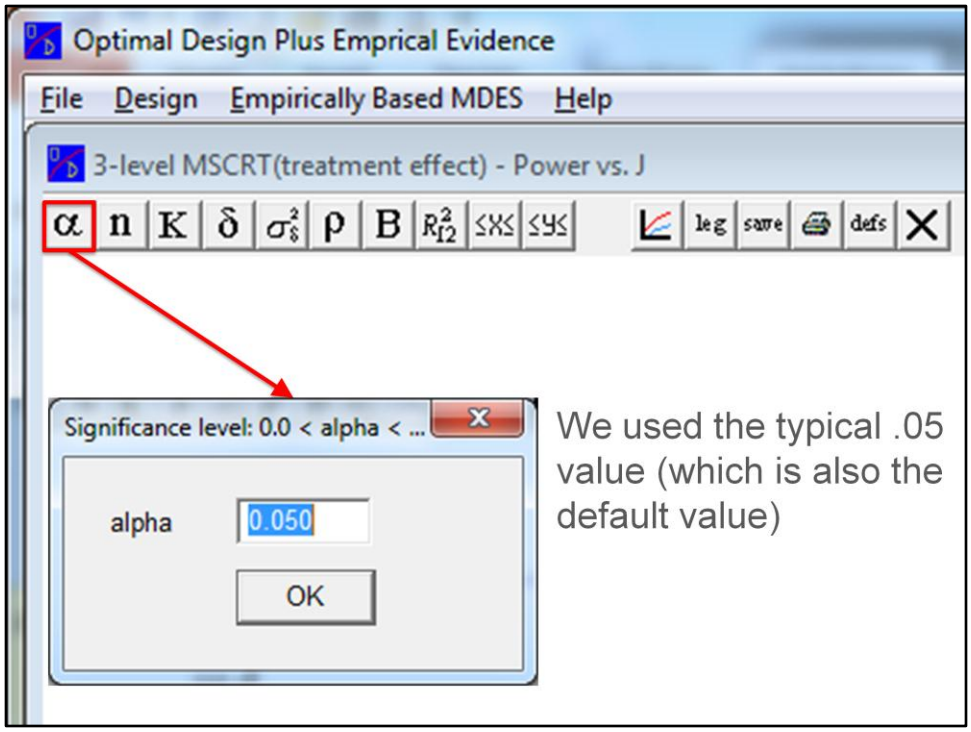
Then, because our intervention is at the classroom level, select “Treatment at level 2”



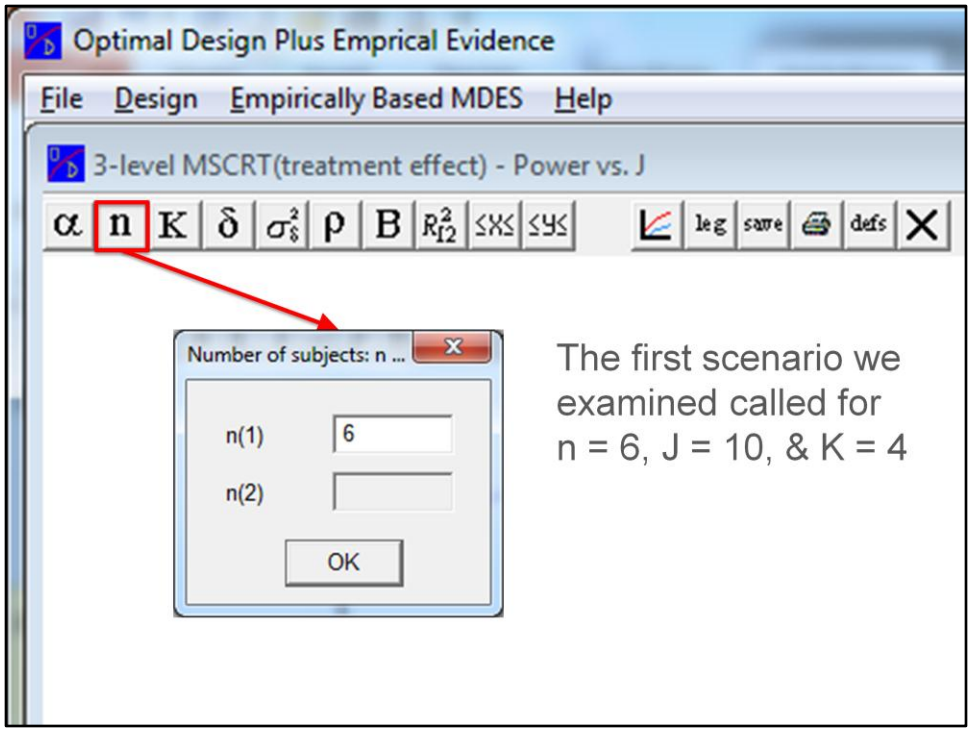
That brings up a menu of options for different power analysis graphs. Let's zoom in on part of that menu.



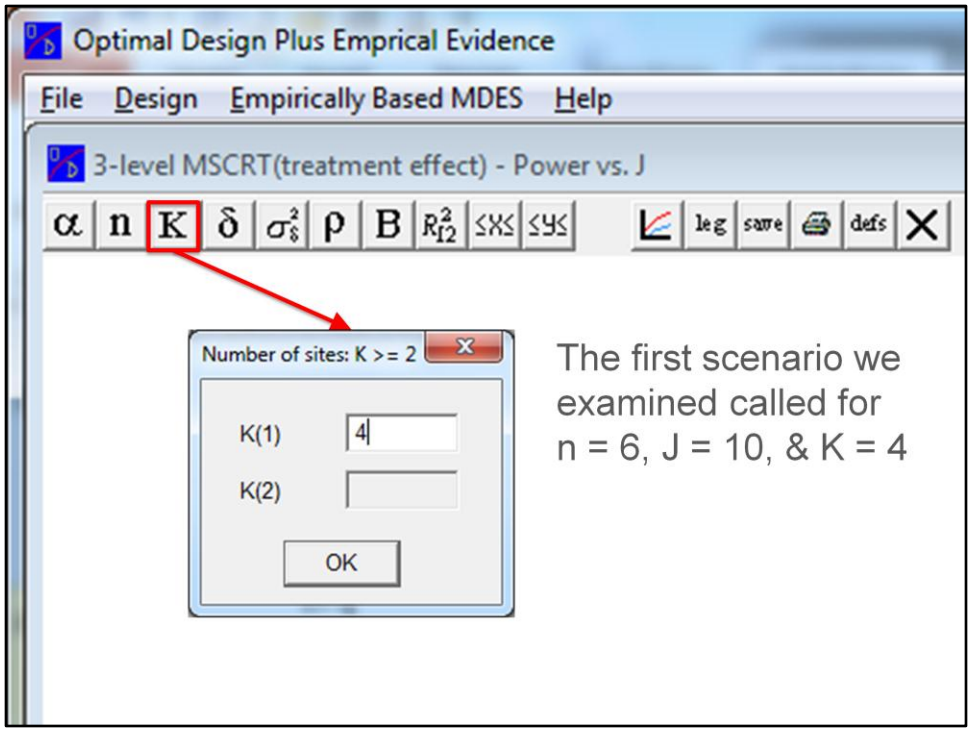
I selected “Power vs cluster/site”. This will allow us to create a graph showing how much power you have as a function of the number of clusters per site (which is the sample size parameter J).



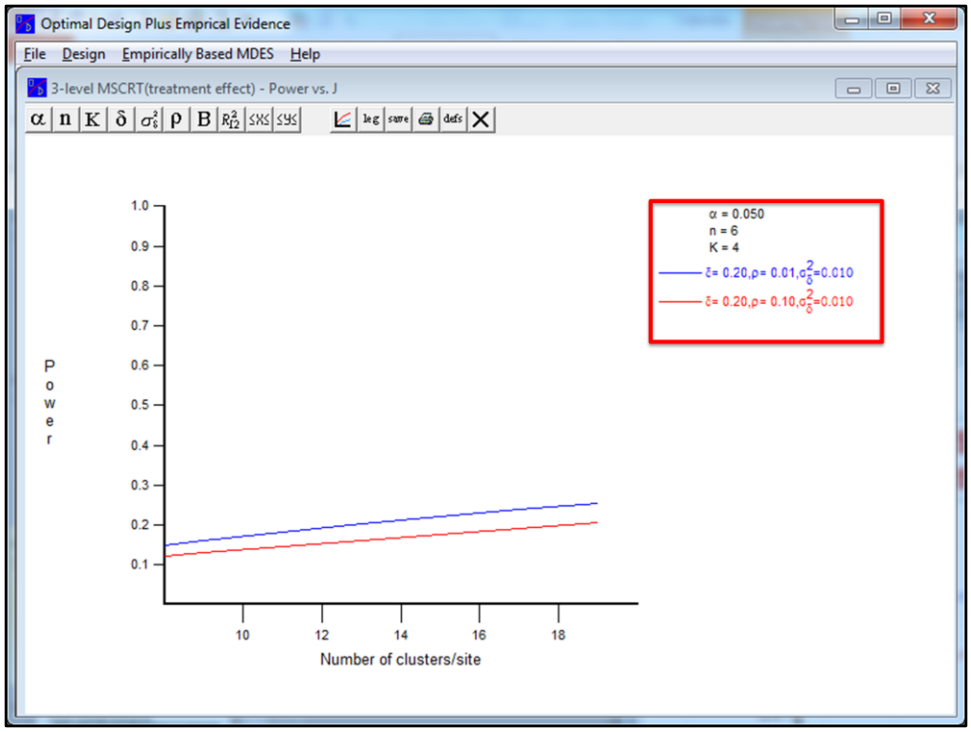
That brings up a new window inside Optimal Design that looks blank except for a row of buttons at the top. To start creating an actual graph, click the button labeled alpha. The result is a little dialog box where you can enter the Type I error rate. I used the conventional value of .05 for all of our power analyses. Clicking OK then creates the first graph.



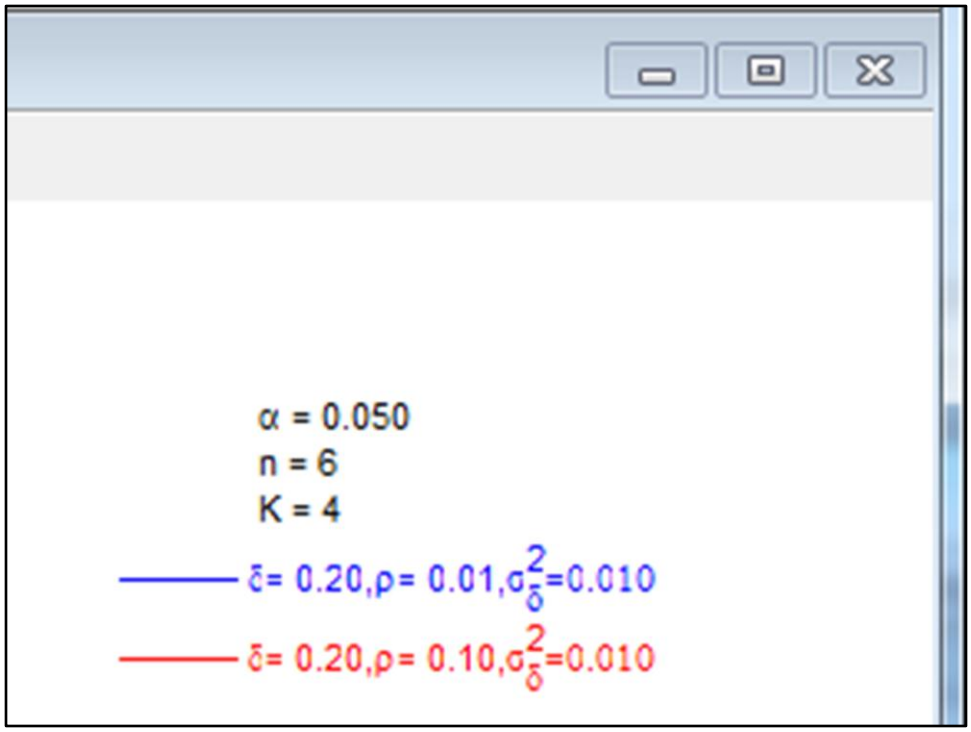
The resulting graph is based on default values for the other input parameters, so we still need to use the other buttons at the top of the screen one by one to set those other parameters to values specific to our study before examining the graph. Since the initial question specified a set of proposed sample sizes at each level, let's fill those in one at a time. I'll start by setting the number of children per classroom by clicking the button labeled n.



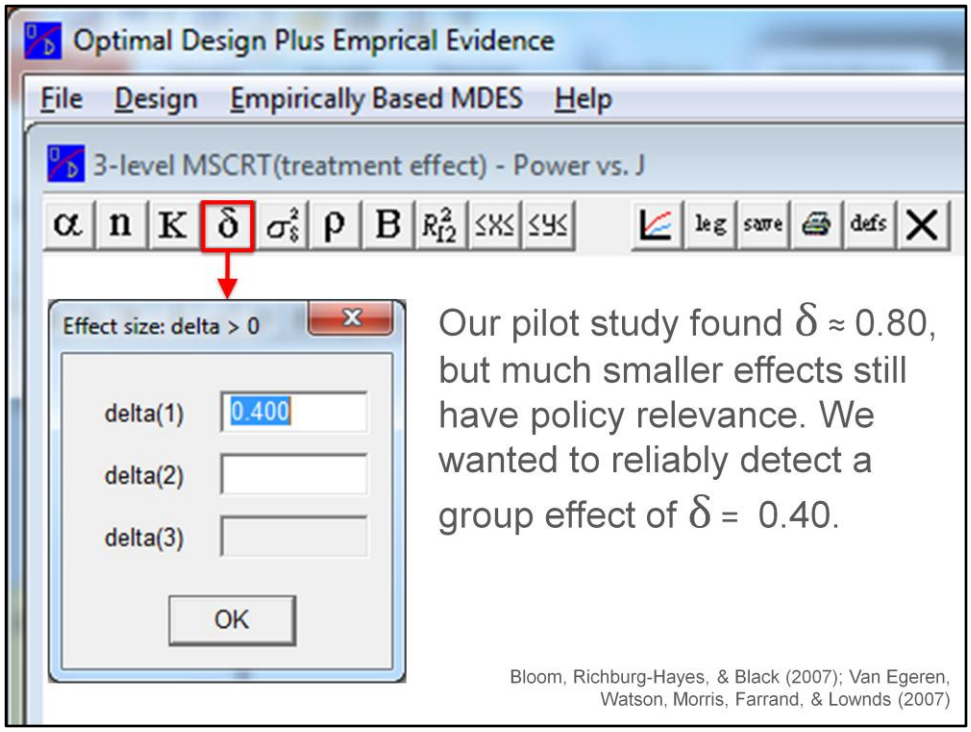
Next, I use the button labeled K to tell the software that we want to have 4 sites. That will update the graph on the screen.



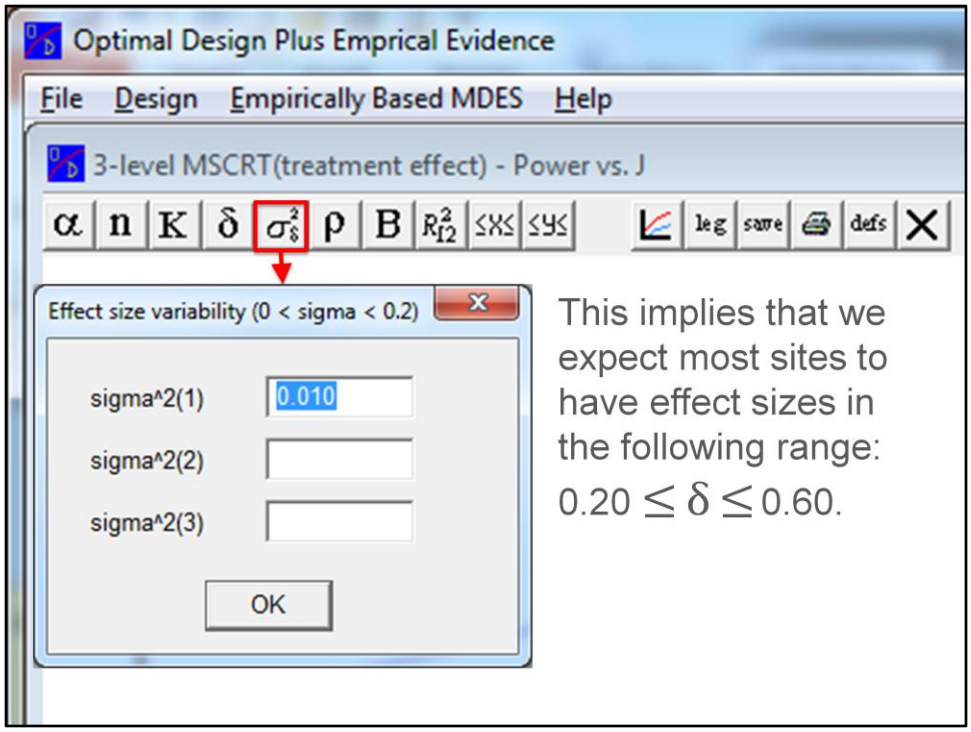
Hmm, so far the power looks awfully low – it is less than 20% for J = 10, but notice the legend in the upper right corner. It's showing the current values used for various parameters. Let's take a closer look at them to see whether they are consistent with what we think is appropriate for our study.



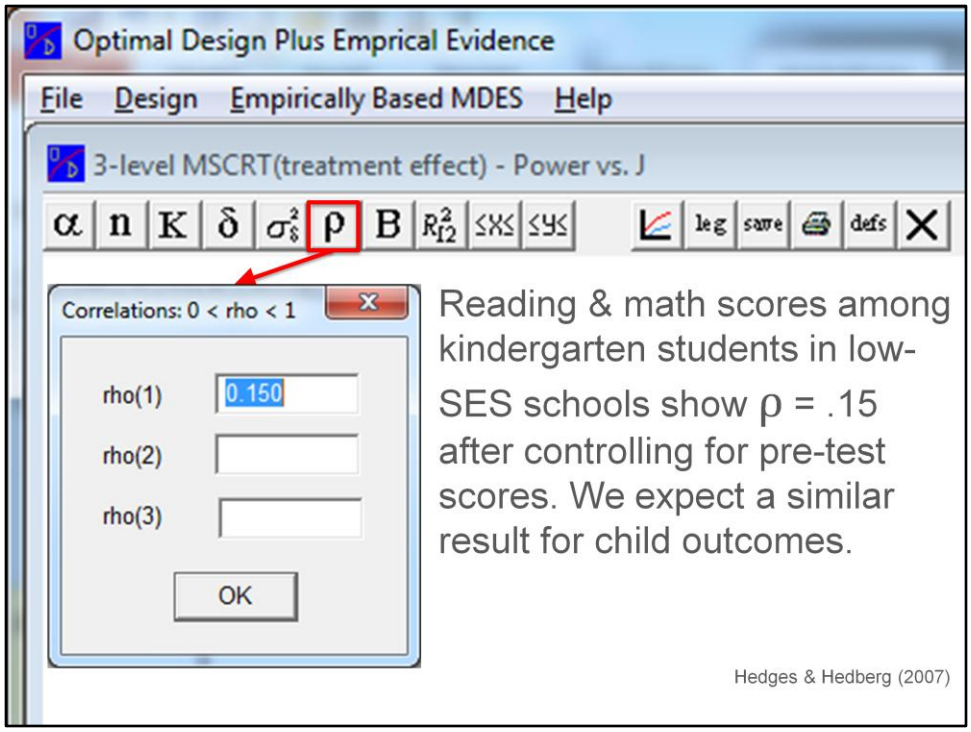
This shows that the graph is currently based on a standardized effect size of 0.20, with an ICC of either .01 for the blue line or .10 for the red line, and a between-site variance of .01 for the effect size for both lines. Those are the default values in the software, so we need to set them to values that make more sense for our study. This is where having pilot data our or previous literature to draw on can be extremely useful. These are parameters that you really want to make informed choices because otherwise your power analysis could lead you very far astray!



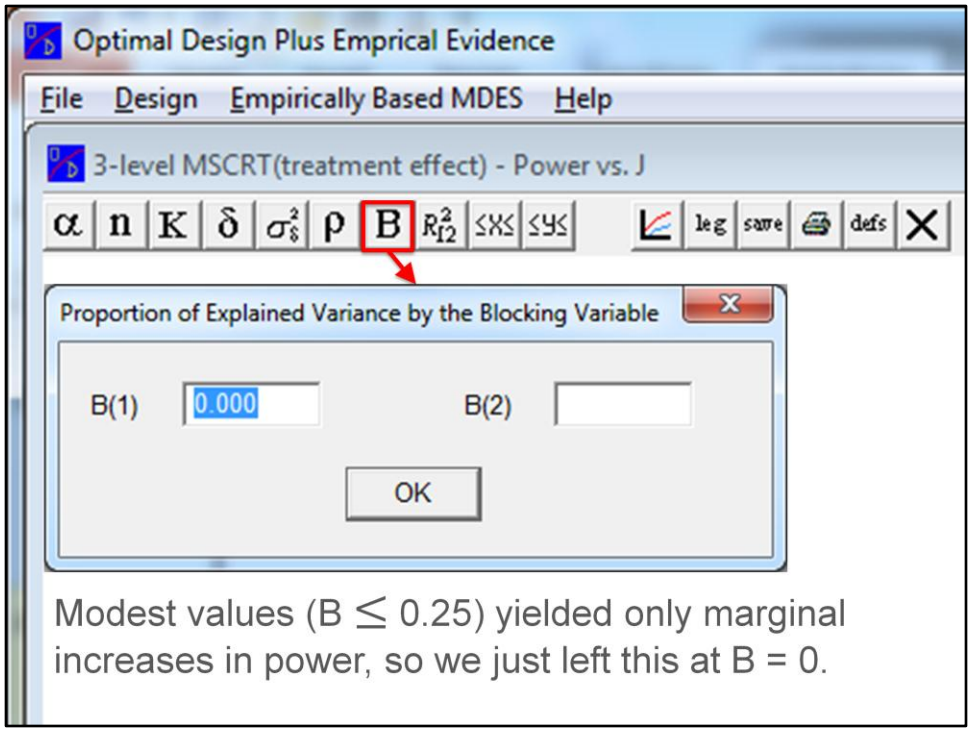
So, next we click on the button labeled delta and insert the effect size we hope to be able to detect. In Laurie's pilot study, the intervention group outperformed the control group by about 0.80 standard deviations. That's a really large effect, so it should be easy to detect with even a small sample. However, Bloom and colleagues have pointed out that even much smaller effect sizes can have very important ramifications for educational policy. Furthermore, there's always the possibility that the pilot study over-estimated the effect size. So, we decided to power the study to detect an effect size of 0.4 standard deviations. We thought this would be a meaningful effect size to achieve, given the general costs of implementing the intervention. If the effect is larger than that, our study will simply have more power than we expect, but using the smaller effect size in planning the study seemed prudent.



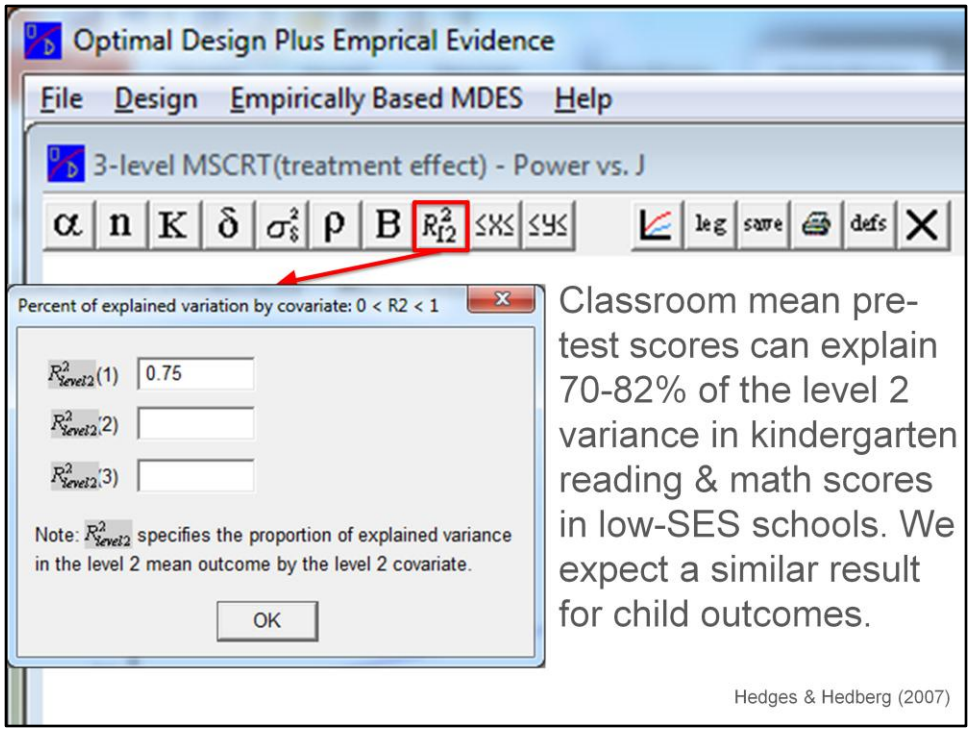
Next, we set the between-site variance in the effect size by clicking the button labeled sigma-squared. We entered a value of .01, which implies a standard deviation of 0.1. That means we are allowing the treatment effect to vary from about .2 to .6 for most sites. So, we expect the intervention group to score better than the control group at nearly all sites, but better at some than at others. This also means we're generalizing our findings to a larger population of Head Start programs than just the ones in our study.



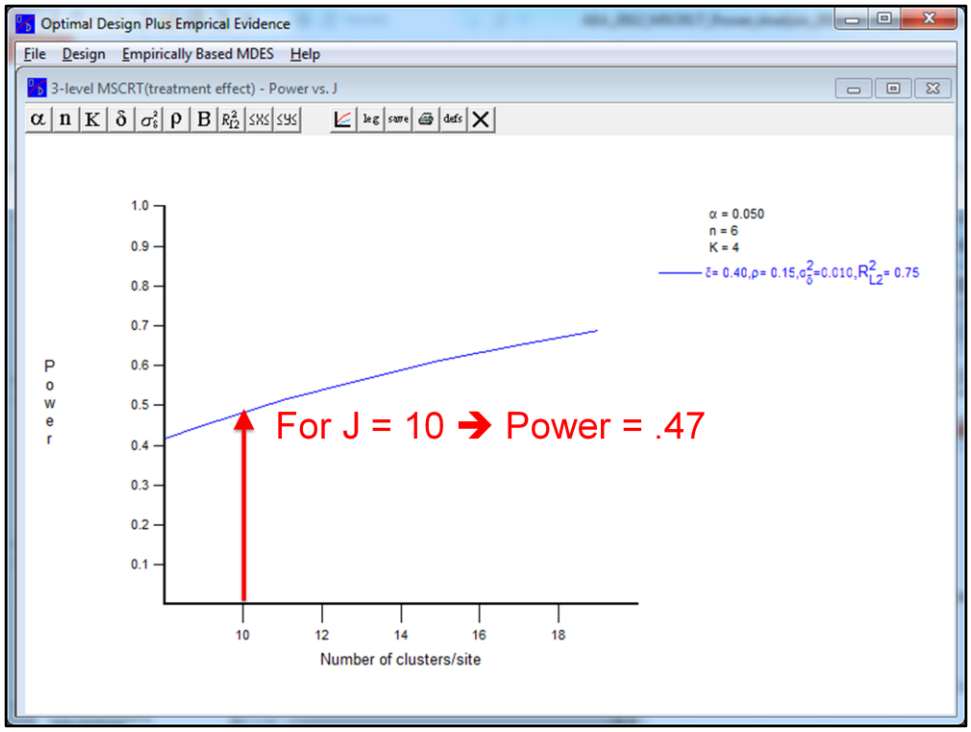
Next, we move on to setting rho, the ICC. Here, I drew on an educational study by Hedges & Hedberg showing that reading and math scores among kindergarten students in low socioeconomic status schools often have ICCs around .20, but that by using the students' pre-test scores as level 1 covariates, you can reduce the ICC to about .15. Technically, this pertains to different outcomes than we will be focusing on, but we expect the result will be similar for our measures of child outcomes, one of which is a measure of scientific reasoning and problem solving. These studies come from large national samples, pertain to kids only slightly older than our sample will, and represent economically disadvantaged populations such as typically use head Start. So, they seemed like a reliable source for these estimates. We planned to use pre-test a covariate, so I plugged in .15 for rho.



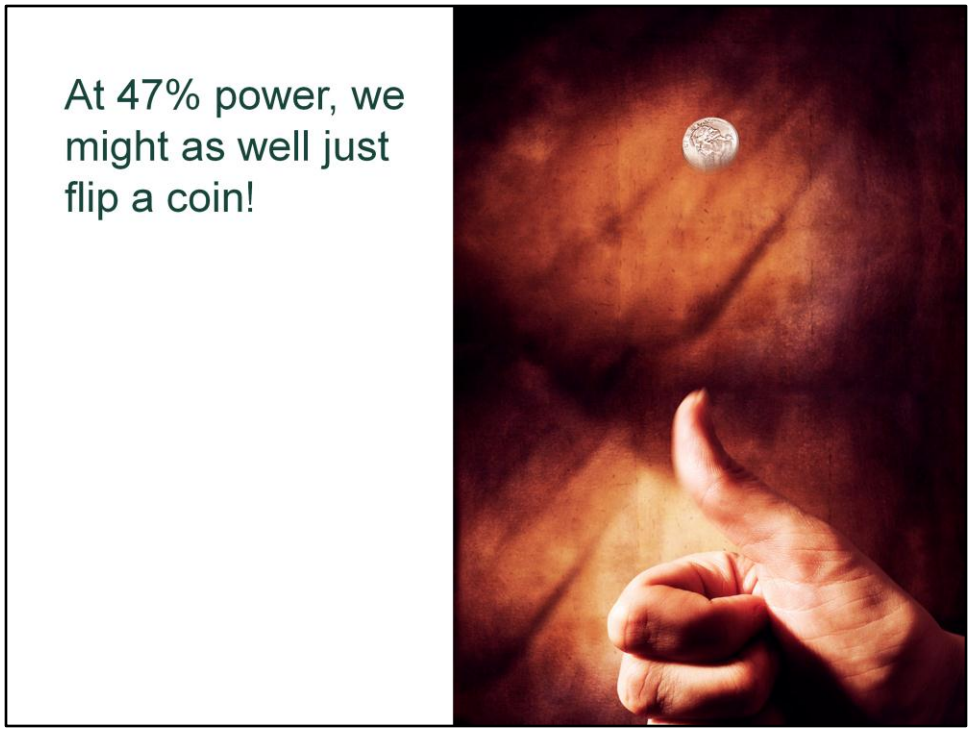
Next, I went on to consider the variance explained by blocking covariates at the site level. Here, I tried a few different values because I thought that our cohort effect might have a mild effect, but I doubted B could really exceed about .25. When I looked at the results, it became clear that the values less than that had only a minimal effect on the power, so I just left this parameter at zero.



Finally, we come to the last parameter: The proportion of variance explained by level 2 covariates associated with classrooms. In our model, this includes both the minimization covariates, plus the classroom-level mean pre-test score. I planned to use that latter covariate because Hedges and Hedberg also reported that such mean pre-test scores can explain between 70 and 82 percent of the level 2 variance in outcomes like reading and math achievement. That has a very beneficial effect on power, so I assumed we could probably see similar results with our child outcomes. I plugged in .75 here.



Ok. So having plugged in values for all of the parameters, let's look at the resulting graph. It shows us that given the sample size Laurie originally asked me about (n = 6, J = 10, and K = 4), we would only have power of .47.



With power at just .47, we might as well just flip a coin instead of collecting data. It's faster, cheaper, and more likely to lead to a correct inference about whether there is a treatment effect.

We need to aim for higher power, at least .80. That's a conventional target for adequate power. So, then we tried some alternate combinations of sample sizes to see what it would take to hit that target.

Power Analysis Scenarios

α	δ	ρ	R^2_{L2}	σ^2_δ	B
.05	.40	.15	.75	.01	.00

n	*J*	*K*	Power
6	10	4	.47
8	10	4	.54
10	10	4	.59

Just increasing *n* doesn't help much.

Unlikely to successfully recruit >10 children per classroom.

We need to adjust *J* & *K*!

Having shown you how to plug in the various parameters, now I'm just going to condense the results of various scenarios into some simple tables to highlight the results of exploring various options. The top table here summarizes the inputs that remained the same across all the scenarios because we had good rationales for those values. I focused on varying sample size at this stage.

As you can see here, if we leave J and K at 10 and 4 respectively, just increasing the number of children per classroom doesn't help power very much. In discussions with Laurie, I determined that we were not really confident that we could expect to successfully recruit more than about 10 kids per classroom, so that is an upper bound on small n. Clearly, if we want more power, we have to try adjusting J and K!

Power Analysis Scenarios

α	δ	ρ	R^2_{L2}	σ^2_δ	B
.05	.40	.15	.75	.01	.00

n	*J*	*K*	Power
6	4	8	.55
8	4	8	.64
10	4	8	.70

Minimums of *J* = 4 & *K* = 8 seem sensible.

Power still not high enough even with *n* = 10.

Try increasing *J* again!

So here is another set of scenarios where I decided to try increasing K to 8, which seems like a lower bound on the number of sites given the literature on estimating variance components for the top level in a model. However, just to see what would happen to power, I dropped J all the way down to 4, then varied n from 6 to 10.

This improved power compared to the original scenario, even though we'd actually be using 32 classroom instead of the 40 previously planned. Still, it didn't quite get us where we wanted to be on power. So, next I tried increasing J again.

Power Analysis Scenarios

α	δ	ρ	R^2_{L2}	σ^2_δ	B
.05	.40	.15	.75	.01	.00

n	*J*	*K*	Power
6	4	8	.55
6	6	8	.73
6	8	8	.84

Increasing *J* to 8 solves the problem!

Here's another set of scenarios. This time, both small n and K are fixed at 6 and 8 respectively, and I increased J from 4 to 8. wanted to keep J in even multiples of 2 so that we could have equal numbers of intervention and control classrooms at each site. That last scenario looked really good – power of .84 is really respectable.

Oversample to Protect Against Attrition

Required	Required *N*	Attrition	Recruit	Recruit *N*
K = 8	8 programs	0.00	K = 8	8
J = 8	64 classrooms	0.25	*J* = 10	80
n = 6	384 children	0.25	*n* = 8	640

Required * (1 + Attrition) = Recruit.

Round result up to next integer.

The next step was to calculate how much we needed to oversample in order to still end up with those required sample sizes after any losses to dropout and attrition. We estimated that there would be 25% attrition at both the classroom and child levels, so I inflated the required sample sizes accordingly and determined that we really needed to recruit a total of 8 programs, 80 classrooms, and 640 children in order to end up with complete data for 8 programs, 64 classrooms, and 384 children by the end of the study. I sent those numbers off to Laurie and she inserted them into the grant proposal.

Proposal Feedback From NSF

Proposal Feedback: Budget Concerns

Positive scientific review comments
NSF very interested, but…
$2,918,640 is too costly
Cut to about $ 2,500,000

When we got the feedback from NSF, we learned that the reviewers scored us very well and gave very positive scientific review comments. NSF was very interested in the study, but … felt that the $2.9 million budget was too costly. So, NSF asked if we could cut the budget to something closer to $2.5 million.

Our team asked how we could cut the cost that much without compromising the study. Laurie noted that our largest costs were in data collection. I suggested maybe we could collect less data without losing too much power and offered to check some more power analysis scenarios to see if we could find an acceptable tradeoff of power, sample size, and cost.

Revised Power Analysis

Revised Power Analysis Scenarios

α	δ	ρ	R^2_{L2}	σ^2_δ	B
.05	.40	.15	.75	.01	.00

n	*J*	*K*	Power
6	8	8	.84
6	7	8	.80
5	7	8	.74
6	6	8	.73

Decreasing *J* to 7 causes minimal loss in power.

Group sizes may not end up equal within each site.

So here are the 4 scenarios that I tried. The top row is what we proposed originally, the others were new scenarios. Laurie had suggested maybe reducing total classrooms would do the trick and indeed just dropping the required number of classrooms from 8 to 7 still preserved power of .80, while allowing her to trim the budget substantially. It did mean that groups sizes may not be equal within each site, but that was an acceptable sacrifice.

Oversample to Protect Against Attrition

Required	Required *N*	Attrition	Recruit	Recruit *N*
K = 8	8 programs	0.00	*K* = 8	8
J = 7	56 classrooms	0.25	*J* = 9	72
n = 6	384 children	0.25	*n* = 8	576

8 fewer classrooms

64 fewer children

Here's the final implications of the revised power analysis for how much we needed to oversample. I used the same attrition rates as previously, but now we needed 8 fewer classrooms and 64 fewer children overall.

Revised Budget

We cut $ 325,641 (11%) partly by reducing *N*

Power only dropped from .84 to .80

NSF accepted $2,592,999 budget

Ultimately, we were able to cut over $325,000 from the budget – about 11% – and a good chunk of that came from the reduced sample size. We only sacrificed 4% power along the way.

NSF accepted our revised budget of about $2.6 million. So, thoughtful application of power analysis helped us land a large grant.

Use power analysis to plan sample size in your own work!

My take-home message for you is simple. I hope you will use power analysis to plan the sample sizes in your own work. I hope this example showed you some of the principles involved in using these sorts of techniques.

MICHIGAN STATE UNIVERSITY

Slides will be posted to the AEA eLibrary
http://comm.eval.org/

Steven J. Pierce pierces1@msu.edu
Laurie Van Egeren vanegere@msu.edu
David Reyes-Gastelum reyesgas@msu.edu