00:00:00.000 --> 00:00:01.197 So welcome everyone,
00:00:01.197 --> 00:00:03.990 it is my great pleasure to introduce
00:00:04.067 --> 00:00:05.887 our seminar speaker today ,
00:00:05.890 --> 00:00:07.279 Doctor Elizabeth Tipton.
00:00:07.279 --> 00:00:09.131 She’s an associate professor
00:00:09.131 --> 00:00:11.846 and practice Center and a faculty
00:00:11.846 --> 00:00:13.936 fellow in the Institute for Policy
00:00:13.936 --> 00:00:16.274 research focuses on
00:00:16.274 --> 00:00:18.428 the design and analysis of randomized
00:00:18.430 --> 00:00:20.278 experiments with a focus on issues for
00:00:20.278 --> 00:00:22.588 external validity and generalizability,
as well as meta analysis with the focus on dependent effect sizes.

Um, today she’s going to share with us how to design randomized experiments to better understand treatment effects.

Head virginity welcome best.

The floor is yours.

Thank you. I’m very excited to be here today.

I really wish I hear I wasn’t talking about my office slash closet and was actually with you guys in person and this is my first time doing slides where I’m on the slide so it’s a little.
I don’t know what is the protocol for questions. How do you guys usually set this up? Do people what’s the norm? Do you guys usually jump in with questions or save them for the end? So I think as you prefer, we can do either way. OK, I’m just I won’t be very good at checking the chat, so if there’s a question if somebody can just speak up that would be will do that. I’ll do that on the chat. OK, thank you. So I just want to set out background for what I’m talking about today, which is I’m talking about randomized
I realized that in a Biostatistics Department, you guys. The idea that randomized trials are common is probably almost absurdly basic for the world that you operate in, but I do a lot of my statistical work in the areas of education and psychology and kind of in the field experiments world and those areas. Randomized trials have only become common really, I’d say the last 20 years. So almost 20 years ago, the Institute for Education Sciences was founded in the Department of
Education in the US government, and that has funded almost 500 what are called efficacy and effectiveness trials and education. Previous to that there were very few of these. There’s also an increasing number of nudge experiments in social psychology experiments that are occurring in the world. I know that there’s a lot of rain in mice trials occurring in developing countries, so this is late in parallel, maybe 2. In public health, they’re being randomized trials there, so I’m just sort of pointing out
that these are becoming increasingly common for policy decisions, not just individual decisions. But the trials as they are designed currently are not necessarily ideal in the sense that they are not as big as we would like them to be. In order to be able to really explore the data well, there often in, not as big as we would like them to be. In order to be able to really explore the data well, there often in, not as big as we would like them to be. In order to be able to really explore the data well, there often in, sort of somewhat small samples in the in the kind of education world that I work in it. They’re very often just simple to arm designs 5050 treatment control.
I much less common to see things like step wedge or smart designs, so those are trickling in, I think. And the goal of these is often to get into some things like clearing House of some places so that policy making decision makers can use the information from the trials to make decisions. But the problem which is the focus of my talk is that there very often been taking place in samples that are purely of convenience, which makes thinking about generalizability and heterogeneity rather difficult. If the treatment of X very if treatment
effects vary across individuals or they vary across clusters in some way, then it’s pretty straightforward to see as a group of statisticians here that the average treatment effect you would get in the population. Is probably not exactly the same thing as the average treatment effect in the sample, and that these could be quite different if treatment effects vary a lot and depending upon how the sample is selected. So there has been an increasing amount of work in this area. There’s a couple of papers I think that are particularly helpful if there’s
a paper in education where they’re looking at bias from non random treats. Non random treatment assignment or they show that the bias of external validity is on the same order as internal validity bias and so to do so they hear they sort of leverage natural experiment with a randomized trial to look at this. And that’s worked by Bell, Olson, Oregon, Stewart. There’s also work showing that. In education and the kinds of schools and school districts that take part in randomized trials are different than the populations of various populations.
At something like the Institute of Education Sciences might be interested in, so I have a paper out with Jessica Spy, Brooke Ann are students looking at 37 randomized trials and the samples of schools taking part in those studies and comparing them to various populations of schools. In the US. There’s also work hidden behind me. By Liz Stewart. San colleagues looking at school districts and a couple of other papers as well, and these find fairly consistent things. For example,
that large school districts are overrepresented in research. Relative to the size of districts in the US. There's been also a lot of work in this area of generalizability and post hoc corrections. I started into this work looking at using post stratification as a way of estimating a population average treatment effect from a sample. There's also been work using inverse probability weighting, maximum entropy weighting bounding approaches. There have been some approaches that focus on little,
so I’m thinking like.

Here’s the paper and Stuart San Green.

I think that does that, so there’s been like a kind of a flurry of method development.

I think here in this area of thinking you know, how do I actually estimate this?

If I have population data of different forms and I have sample data of different forms, how can I actually estimate a population average treatment effect?

But when I first started doing this work, I realized in a series of examples that I was working on that the
effectiveness of these methods is often severely limited in practice because of undercoverage and what I mean is that it you can’t. If it turns out that your population has there’s a part of the population that’s just not represented in the trial, there’s really not much statistical magic you can do. You can make some assumptions, but you can’t really re wait something that doesn’t exist, and the. It’s it’s really a reflection of lack of positive ITI in the study. Yes, exactly thanks. Yeah exactly. And yeah, I’m just using survey
sampling language for the same thing.

That’s right, yeah, and so.

And that’s the lack of positive ITI

often arises because people aren’t

thinking about what the population is

in advanced and so it’s very tricky

for them after the fact to generalize,

because it turns out that maybe

this what I as analyst him now

trying to think of as the population

isn’t exactly the population,

but it’s very hard for people to

articulate what the population is,

and so spent a lot of time just trying to.

You’re out what the population actually is,
and if that’s population is meaningful, as if that’s a population that even matters.

So I realized I pivoted a bit. I realized that you could do a lot of this with statistics, but you were going to be limited if you didn’t design better trials. And so that’s allowed me to think. Well, why don’t we just start? The beginning and do this do a better job this so why?

What have we started at the beginning of our studies by asking what the target population of the intervention was thinking about inclusion and exclusion criteria,
I think it helped.

This probably matters even more with like comorbidities and you know.

I like rolling out people.

You’re doing a study on depression, but you rule out people with anxiety, and that’s like a big problem for the interpretation since they were highly related to each other.

This is true in education as well.

So like, what are the inclusion exclusion criteria for your trial and how might that affect where you can generalize? And then also thinking about
background characteristics and
contextual variables that might moderate the intervention’s effect
and the tricky part here is, there’s a little bit of a circularity
which I’m going to keep coming back to in what I’m talking about,
which is in order to know these, you know,
we don’t know what these are in advance, and we don’t have a lot of
knowledge generated to date about
because studies have not been designed to estimate or test hypothesis about moderation,
and so instead we have to sort of think
00:09:19.294 --> 00:09:21.357 through what we think might matter.

00:09:21.360 --> 00:09:21.707 Using,

00:09:22.401 --> 00:09:24.830 you know, not a great source of knowledge here,

00:09:26.910 --> 00:09:28.788 but the idea is that you sort of take all of this information and

00:09:31.241 --> 00:09:33.461 then you use this to create to

00:09:33.461 --> 00:09:34.793 use sampling methods to actually

00:09:34.793 --> 00:09:36.245 like using stratified sampling.

00:09:36.245 --> 00:09:37.920 Figuring out if you should

00:09:37.920 --> 00:09:39.400 you know within Strata,

00:09:39.400 --> 00:09:41.360 using balanced sampling or random

00:09:41.360 --> 00:09:43.677 sampling and thinking about sort of

00:09:43.677 --> 00:09:45.714 ways in which you can increase the
coverage so you can have positive ITI for the whole target population, so that when you do you know. When you do need to make adjustments at the end of your trial using these statistical methods, they are in a realm in which they can perform well. This sort of led me to thinking about tools for generalization, and so I just want to highlight this because I think this is a good strategy for methods people to think about. So I thought will nobody is going to do what I’m telling them to do if I don’t build a tool because the kind of people.
Clan randomized trials, at least in my domain, don’t often have statisticians ready at the ready to work with them on things, and they are often writing grant proposals before they’ve got funding, and so it’s very possible that they’re not going to think about generalization or or have the training or tools to do it so. I got a grant from the Spencer Foundation and then I’ve had follow up money from the Institute of Education Sciences to build this tool called the Generalize are that uses some basic design principles standalone.
It’s got.

It’s very focused on the user experience and it in the background has the Common Core of data which is.

An annual census of the public schools in the US so that the data is already been cleaned and set up.

We’re adding in right now the iPads data, which is higher Ed data in the US and so the idea is somebody could go in and walk through inclusion exclusion criteria identified moderate or.

It would build you stratified recruitment plan in less than an hour.

You could leave with a list of all the schools and start being able
to recruit with the.

Great, I've had this going since 2015 and it was very slow going for awhile. This is sort of a year that I could actually extract a lot of user data and So what you can see here is actually it was slow going and I had some early adopters. These are people that would be star users, so many of them are planning randomized trials, but there was actually a very big jump that occur this summer and that’s based on this jump. I actually started digging through...
things and realized that.

Institute of Education Sciences that actually enacted requirements for generalizability.

In their request for proposals, and so you can see that what I already always speculated, which is that funders really drive change. So once funders said you need to pay attention to generalizability, people actually started paying attention to generalizability in their proposals.

OK, so this I just wanted to give you all of this background as a way of explaining sort of my like where I’m coming from in the in heterogeneity and how I’m
NOTE Confidence: 0.83171284
00:12:35.925 --> 00:12:36.984 thinking about this.
NOTE Confidence: 0.83171284
00:12:36.990 --> 00:12:37.391 Um?
NOTE Confidence: 0.83171284
00:12:37.391 --> 00:12:39.396 So everything I’ve talked about
NOTE Confidence: 0.83171284
00:12:39.396 --> 00:12:41.791 is sort of averaging over hedge
NOTE Confidence: 0.83171284
00:12:41.791 --> 00:12:44.297 and 80 when we talk about analyze.
NOTE Confidence: 0.83171284
00:12:44.300 --> 00:12:45.648 Estimate an average treatment
NOTE Confidence: 0.83171284
00:12:45.648 --> 00:12:46.996 effect for a population,
NOTE Confidence: 0.83171284
00:12:47.000 --> 00:12:49.060 assuming that there’s variation of
NOTE Confidence: 0.83171284
00:12:49.060 --> 00:12:51.560 effects and we’re averaging over those.
NOTE Confidence: 0.83171284
00:12:51.560 --> 00:12:53.576 But to average over those requires
NOTE Confidence: 0.83171284
00:12:53.576 --> 00:12:55.686 that we know something about how
NOTE Confidence: 0.83171284
00:12:55.686 --> 00:12:57.804 treatment affects very and very often,
NOTE Confidence: 0.83171284
00:12:57.810 --> 00:13:00.085 and I would say this is the
NOTE Confidence: 0.83171284
00:13:00.085 --> 00:13:01.620 in general we don’t,
NOTE Confidence: 0.83171284
00:13:01.620 --> 00:13:04.100 and the reason that we don’t have a
NOTE Confidence: 0.83171284
great handle on this is because sample size and sample sizes in randomized trials.

I’ve been very focused on the after treatment effect.

Moderators have only become more of a focus, at least in education.

More recently, and I think that’s true in psychology and related areas as well. And they are often more like exploratory analysis at the end, so you end up with these problems where moderator effects don’t get replicated and they don’t get replicated because there was.

You know,
who knows how many statistical tests conducted in order to find those moderators. So they’re not very stable, and we don’t really necessarily understand or their underpowered deeply underpowered like you just have a very homogeneous sample. And so how are you going to find a treatment effect variation if there’s not much variation in your sample to start with, so they’re often an afterthought, but I what I noticed overtime is that as generalizability has become something people are paying attention to,
people are also starting to pay attention to the idea that you could predict treatment effects, or that you could identify subgroup effects and that this might be very useful information. Which led me to start thinking about how you would design trials for this. So what I’m going to, what I’m leading up to is talking about designing trials to think about heterogeneity. So I’m just going to start with like a little bit of a background here. So we’re going to assume that you’ve got.
I’m assuming we’ve got units which are usually here. Let’s say students insights which might be schools, and I’m doing a randomized trial, and I’ve got these potential outcomes. And so we’ve got both an average and intercept in these, and we’ve also got some sort of fixed variation that we can explain. And then we have this other parts that are not affected by the treatment. We’ve got some site level and individual residuals and some idiosyncratic errors. But what we’re interested in
really is in these these moderate
yrs of treatment effects,
and so you could say that Delta
0 is the difference in averages.
I’m assuming these are centered variables,
so this is nicely the difference in
averages and that the vector Delta
is the difference between these
effects of the treatment and then
under treatment and under control.
A lot of so as you have to
think about interpretability
here of what I mean by Delta.
By this by these deltas and
how to standardize because we
wanted to talk about these,
they need to have a mean of 0, but also in order to talk about treatment effects. 
Sort of done in general for developing things like power, we often standardize them so often we have effect sizes for the average treatment effect, their standardized in relation to the variation in the in the sample and the population. And so here I’m going to sort of say we what we need to do is we need to standardize the covariates and we need to standardize the
covariates in relation to the population standard deviation. This might not seem like this is like a radical statement, but if you look into the power analysis literature on how to conduct power analysis for moderate are tests, they are typically standardizing in relation to the sample standard deviation, and in doing so, it makes it impossible to see how your sample actually how you choose your sample might matter. Isibaya standardizing by this fixed value by the population, you’ve identified a population,
and now we’re standardizing by that population standard deviation. That will make the role that the sample plays here much more clear.

So the fact that we randomized to treatment and control allows us to estimate these dealt these Spectre Delta using some generalized least squares of some sort, and I’m being a little big here because I’m trying to encapsulate all like versions of this.
OK, so I can do so, I can separate. These are at additive or rather subtractive. I guess the treatment and the control their step. You can separate them. And and through this I can think about statistical power, and for each of these moderator effects. And so, one way you can do that is through the minimum detectable effect size difference. I don’t know how common this is used in the sort of Biostats world, but it’s a pretty common metric that’s used in cluster randomized trials in.
The world I work in, and so it’s nice because it’s sort of easily interpretable,
so this is the smallest affect size that you could for a given Alpha level which is affecting this.
That’s like the critical value.
This is sort of the smallest true effect that you could detect with the power that you with like
80% power for example.
And so this is like a general form for this,
and So what I’m showing is that its function,
can I like move my hands?
I don’t know.

I’m just going to involve a lot of never mind,

so it’s a function of the variation in the population in that covariate.

It’s also a function of $S$, which is you could think of as the sort of covariance matrix of the $X$ is in the sample,

And then it’s a function of $N$, and then it’s constant here.

which is the sample size per cluster. I’m assuming it’s constant here.

$J$ is the number of clusters and $P$ is the proportion in treatment. So

that what is $\sigma_{XK}^2$?
The population SD of effect modifier or the population variance effect modifier? Is the population variance of the effect moderate or modifier? But then your square rooting it so it’s going to be gradual scale.

OK, so just to give you a couple of special cases where you can sort of parse out some things. So there’s been previous work. I meant to include a citation here by Jessica Spy, Brooke and colleagues. That’s looking at power for moderate are tests. And so here’s 2 cases we have.
Site lab, site level, moderate yrs and individual level moderate yrs and.

I’m taking basically what they’ve got, but re tweaking part of it.

Because I’m factoring out this Sigma squared and noting that you can actually pull out this thing called RXK at the front and the RXK is this ratio of the standard deviation of the covariate in the sample compared to the standard deviation of the covariate in the population.

And So what you can see here is by doing that you by rewriting it. This way you can see that our XK is
00:19:45.426 --> 00:19:48.293 having just as much of an effect on
00:19:48.293 --> 00:19:50.433 statistical power as things like the
00:19:50.433 --> 00:19:53.682 square root of N or the square root of P.
00:19:53.682 --> 00:19:55.506 These other parameters that most power
00:19:55.506 --> 00:19:57.178 analysis has spent has focused on,
00:19:57.180 --> 00:19:58.008 and that’s true.
00:19:58.008 --> 00:20:00.260 You know, in any of these designs.
00:20:00.260 --> 00:20:02.836 Love seeing it in any of these designs.
00:20:02.840 --> 00:20:06.114 RX shows up. OK,
00:20:06.114 --> 00:20:10.020 so if RX is something that matters for power,
00:20:10.020 --> 00:20:12.190 a question will be well.
00:20:12.190 --> 00:20:14.360 What are people doing in
00:20:14.360 --> 00:20:16.096 practice right now right?
00:20:16.100 --> 00:20:18.698 So maybe maybe people are choosing
00:20:18.698 --> 00:20:19.997 fairly heterogeneous samples,
and So what I’ve got here is 19.

This is 19 randomized trials in education that we extracted.

So these are box plots of values across each of these 19, and for each of them I’ve calculated for holding like the US population of school.

So this is like the US population moderate are in the sample in these studies to the ratio of that to that.

standard deviation in the population OK, and then I’m looking at boxplots of this and what you can see like do this,
00:21:00.400 --> 00:21:00.737 don’t?

00:21:00.737 --> 00:21:03.433 OK, what you can see here is that

00:21:03.433 --> 00:21:05.230 the bar at the bottom.

00:21:05.230 --> 00:21:08.020 Can you see my cursor?

00:21:08.020 --> 00:21:10.369 Can’t tell if you guys can see my curse.

00:21:10.370 --> 00:21:11.936 No,

00:21:11.940 --> 00:21:12.182 OK, so the bar at the bottom there’s an R

00:21:12.182 --> 00:21:14.360 X = sqrt 1/2 and then there’s a line

00:21:14.425 --> 00:21:16.711 across the top that’s like a dashed one.


00:21:18.980 --> 00:21:20.570 That’s the R X = sqrt 2.

00:21:21.570 --> 00:21:21.885 OK,

00:21:21.885 --> 00:21:24.090 and so you can see that most

00:21:24.090 --> 00:21:26.298 studies are actually below there,
less heterogeneous than the population there.

Below this line for one, and they’re actually far less heterogeneous than the population there are. Actually.

If you look at these median values, many of them are closed 2.5, so they are about 1/4 of the variation as we’re seeing in the population.

So this gives you a sense that if. That there’s, uh, an opportunity to improve, right? Like I could increase power not just by increasing my sample size or increasing.

My sample size in schools or the number of
schools which are pretty expensive,

but I could also increase my power by

changing the kinds of samples that I select.

And so that’s where these numbers came from.

They should have gone to

slightly different order.

So the main point is that design

sensitivity, the way we think,

whether that statistical power or

standard errors or whatever framework

that there this is proportional in

some way to this RX value that we can

improve our design sensitivity by

choosing a more heterogeneous sample.

And so funny, I must have like put
this in here twice on accident.

So this is the same thing but with a line through it.

OK so if once you have that insight that heterogeneity matters,

that it’s actually something that we can include in our power analysis and that is something that is not actually happening in practice.

Then we can start thinking about how we might plan studies differently.

OK, so if.

So how can we improve statistical power?

Well,

a lot of the literature as I was saying,

is focused on improving power by
increasing sample size or instead.

But what I'm arguing here is that you could increase instead this ratio.

You could increase the variation in your sample choosing more heterogeneous sample annual have more statistical power for test of heterogeneity of moderators, and so what would you do with this?

It would mean you know purposefully choosing sites that were more extreme, it might end, and that's easy enough to do in one variable.

And I'm going to talk a little bit about how to do that with multiple variables.
So with a simple, let’s just say we had one single continuous. Moderate are like this is a normal distance normally distributed. This theory would tell us that we should choose half of our sample. We would choose half of our sample from the upper and lower tails and choosing them from the upper and lower tails were actually getting an RX of $\sqrt{2}$. This is actually a rather large, so this is going to create a much more homogeneous heterogeneous sample, thus increasing our statistical power because it’s more heterogeneous than the.
In the population.

Similarly, if we had two correlated normal variables,

when we this is, you know,

we could imagine getting the corners of this.

These are all principles, by the way,

straight up from experimental design.

If you think if you think about it,

there are principles from like 2.

You know two factor studies or

multi factor studies where you’re manipulating and instead I’m just saying

instead of manipulating these factors

were now measuring these factors.

Someplace you could choose them
to be extreme design points.

It gets a little harder once things become correlated,

so when they become correlated,

I don’t have as much sample available to me because there’s just fewer population units in those corners,

and so it’s going to become increasingly hard as I add variables,

it might become harder and harder in order to figure out what these units are that I could be sampling from.

So I started thinking about how you would do this,

and I realized that there is actually a literature on this in the world.
sort of industrial experiments and industrial experiments, and in psychology people again are thinking about multi factor studies. So they’re thinking about things you could better in the experimenters control. But we could instead bout sampling in the same as the same kind of thing. Except that we don’t have control over manipulating them. We can find these units and as an alternative approach, so one of the things we want to do is we want to make sure that we observe the full range of
covariate values in the population, so it requires us to actually think, you know, explore the population data and make sure that we can understand what that range of values is. We might need to think carefully about moderators that are highly correlated. It can be very hard to alias these effects, so if you have two highly correlated moderators. I think about that. I have two highly correlated moderators like this. If I want to estimate and understand moderators of X, if I want to explore X&Z and
these are highly correlated, I’m going to really need to make sure I have those off diagonals that are kind of more rare in order to help me separate these effects and understand the unique contribution of each. The other is that if we might have many potential moderators that we’re interested in, we might need to make here. But also think very carefully, like we’re not going to be able to expand
this study to have a much bigger sample.

So a lot of what I'm trying to operate under the constraint here is,

let's not change the sample size if we don't change the sample size,

but we instead change the height types of units in our study,

how much better can we do?

This is work from the 40s and 60 Forties,

This is work from the 40s and 60s.

A lot of work here by Walt Kiefer,

and a lot of people in
Industrial experiments.
The idea is that you can instead focus on the generalized variance an you want to minimize the generalized variance, which is the determinant. So D is for determinant. And so the design that meets this criteria is one that also conveniently minimizes the maximum variance of any predicted outcome based upon these covariates. So this is great if what you’re headed for is trying to make predict individual treatment effects or site specific treatment effects.
The nice thing about a method that’s been around for a while is that there’s been algorithms developed for doing this. Better out Federov win algorithm is widely used and variations of it and that these are package that there are like statistics package already available that do this.

So in our there’s something called the ALG design package that is set up to actually work through this. Designs that we know are optimal. In other contexts.

You know like our designs like Latin squares, designs etc all become special cases of this.
So this is a much more general framework that doesn’t require as many assumptions. OK, so once you start down this path you realize too that there are some tradeoffs here, so we have. You can easily imagine that the design that is optimal for an average treatment effect, which might be a representative sample. That sort of like a miniature of the population on covariates is likely not optimal for some of these standardized effect size differences where we might need to oversample in order to estimate, and so there’s another.
Benefit of this approach, which is that you can focus on augmentation approach and what that means is you can actually say using these algorithms better billable sites or already for I’ve already got design run so the language of this is these sites become designed runs. And I need to select 10 more. Meaning population units, what? 10 units can I augment it with that will improve that will make this as D optimal as possible given these constraints, and so instead so we’re thinking of population units as possible design runs and sample as design.
NOTE Confidence: 0.87889427
00:29:33.769 --> 00:29:36.448 runs that we’ve chosen to use.
NOTE Confidence: 0.87889427
00:29:36.450 --> 00:29:39.762 OK, so I’m just going to go through
NOTE Confidence: 0.87889427
00:29:39.762 --> 00:29:42.388 an example to talk about this.
NOTE Confidence: 0.87089115
00:29:45.440 --> 00:29:47.456 Don’t have a ton more slides
NOTE Confidence: 0.87089115
00:29:47.456 --> 00:29:49.490 I should say so success.
NOTE Confidence: 0.87089115
00:29:49.490 --> 00:29:51.330 OK, so here’s an example.
NOTE Confidence: 0.87089115
00:29:51.330 --> 00:29:53.466 The success for all evaluation was
NOTE Confidence: 0.87089115
NOTE Confidence: 0.87089115
00:29:57.220 --> 00:30:00.388 The reason I like to use this example.
NOTE Confidence: 0.87089115
00:30:00.390 --> 00:30:02.476 Is that it’s old enough that strangely,
NOTE Confidence: 0.87089115
00:30:02.480 --> 00:30:04.809 they actually published in their
NOTE Confidence: 0.87089115
00:30:03.875 --> 00:30:05.650 paper a list of schools they
NOTE Confidence: 0.87089115
00:30:05.650 --> 00:30:07.624 actually named the schools in their
NOTE Confidence: 0.87089115
00:30:07.624 --> 00:30:09.059 study and characteristics of them.
NOTE Confidence: 0.87089115

56
I have other data on other studies where people have shared with me the names of the schools involved, but it’s all like I have to keep it secret for the IR reason, so that the fact that this is available makes it easier to use. So what I did is I went back and looked at the Common Core of data I identified based upon the study that they were. In the way that they talked about their study, that Title One elementary schools in the US at that time might be a reasonable population to think that
they were trying to sample for.

Title one schools have at least 40% students on free or reduced lunch and meet a few other characteristics and then they identified in the paper 5 variables that they thought were possible. Moderators that would be really important to include here, so they talked about total school enrollment being a factor, racial and ethnic composition so I'm using that here as the proportion of students that are black and the proportion that are Hispanic.
and SES meaning and a professor but

NOTE Confidence: 0.87089115

proportion at free and reduced lunch.

NOTE Confidence: 0.87089115

And they also talk about Urbanicity

NOTE Confidence: 0.87089115

because they tried to make sure they

NOTE Confidence: 0.87089115

had some urban schools in rural

NOTE Confidence: 0.87089115

schools and some other schools.

NOTE Confidence: 0.87089115

So I should say in previous work of

NOTE Confidence: 0.87089115

mine I’ve used this as an example

NOTE Confidence: 0.87089115

and then this study actually ends

NOTE Confidence: 0.87089115

up being a fairly representative

NOTE Confidence: 0.87089115

sample of the population,

NOTE Confidence: 0.87089115

which is interesting.

NOTE Confidence: 0.87089115

Is it because they had no real way

NOTE Confidence: 0.87089115

of they weren’t doing it totally in

NOTE Confidence: 0.87089115

a way that allowed them to compare

NOTE Confidence: 0.87089115

this or to choose this in a way,
but they did a lot of work to try to be representative, and this is much more representative sample then.

I take the modal study is in this domain. OK, so what I did for for this example as I'm comparing for you the actual sample selected, so it’s always these five moderators. The actual sample selected a representative sample selected. If I instead I use something like stratified random sampling, The optimal sample based upon these
five covariates using this ALG

And So what I would do here as I’d say.

So if I used 36 sites that were selected with random sampling with stratified random sampling and then I reserved five of them that were selected using D optimality and then I would change, you know, the number of those.

So you could see this sort of effect.

You know that augmentation would have and then for each of these I calculated a few different statistics.
So you can see how this works.

So one of them is D.

This measure of the optimality.

And I’m going to show you relative measures because it’s a little easier to see with relative measures.

I’m also including B, which is generalizability index that I developed.

It ranges from zero to one and one means that the sample isn’t exact miniature of the population on these covariates.

0 means they like are completely orthogonal to each other and.

Chip in its the index is highly related.
00:33:13.572 --> 00:33:17.729 to measures of undercoverage and how and
NOTE Confidence: 0.86540186
00:33:17.729 --> 00:33:20.854 the performance of reweighting methods.
NOTE Confidence: 0.86540186
00:33:20.860 --> 00:33:24.380 And then the mean are meaning the ratio
NOTE Confidence: 0.86540186
00:33:24.380 --> 00:33:27.347 between the the ratio between the
NOTE Confidence: 0.86540186
00:33:27.347 --> 00:33:30.347 standard deviation in the sample and
NOTE Confidence: 0.86540186
00:33:30.438 --> 00:33:33.888 population across these five covariates.
NOTE Confidence: 0.86540186
00:33:33.890 --> 00:33:36.170 OK, so this is what we get out of this,
NOTE Confidence: 0.86540186
00:33:36.170 --> 00:33:38.130 and so I just want to talk through this and
NOTE Confidence: 0.86540186
00:33:38.178 --> 00:33:40.047 I’m happy to answer questions if there’s.
NOTE Confidence: 0.86540186
00:33:40.050 --> 00:33:43.354 I know there’s a lot going on here.
NOTE Confidence: 0.86540186
00:33:43.360 --> 00:33:46.690 Really wish I could figure out how to do a.
NOTE Confidence: 0.86540186
00:33:46.690 --> 00:33:48.900 Pointer.
NOTE Confidence: 0.86540186
00:33:48.900 --> 00:33:51.528 I don’t think I can point out that way.
NOTE Confidence: 0.86540186
00:33:51.530 --> 00:33:52.259 OK, so OK.
NOTE Confidence: 0.86540186
00:33:52.259 --> 00:33:54.989 So what I have going on here is the number
NOTE Confidence: 0.86540186
00:33:54.989 --> 00:33:57.660 of sites randomly selected is left to right?
So on the left is the D optimal sample, meaning the whole all 41 sites were actually selected using a D optimal algorithm on the. Right is the ideal for the average treatment effect. We’ve used random sampling to stratified random sampling and just like the sample an in the bar right right there that like right up there. This Gray vertical bar is the actual study values for each of these. OK so you can see the actual study and then what I’ve got are three
00:34:29.057 --> 00:34:31.017 different lines going on here.
NOTE Confidence: 0.86540186
00:34:31.020 --> 00:34:33.176 So one line that’s sloping down in
NOTE Confidence: 0.86540186
00:34:33.176 --> 00:34:36.008 solid is the relative D optimality value,
NOTE Confidence: 0.86540186
00:34:36.010 --> 00:34:37.129 so this is.
NOTE Confidence: 0.86540186
00:34:37.129 --> 00:34:39.740 You know the highest value is if
NOTE Confidence: 0.86540186
00:34:39.826 --> 00:34:42.370 it was a D optimal allocation.
NOTE Confidence: 0.86540186
00:34:42.370 --> 00:34:43.658 This is a ratio,
NOTE Confidence: 0.86540186
00:34:43.658 --> 00:34:46.160 and then I’ve got the B index,
NOTE Confidence: 0.86540186
00:34:46.160 --> 00:34:47.890 which is the generalizability index.
NOTE Confidence: 0.86540186
00:34:47.890 --> 00:34:50.298 Is the other solid line going up,
NOTE Confidence: 0.86540186
00:34:50.300 --> 00:34:51.504 and so, not surprisingly,
NOTE Confidence: 0.86540186
00:34:51.504 --> 00:34:53.009 that’s increasing as we get
NOTE Confidence: 0.86540186
00:34:53.009 --> 00:34:54.439 to stratified sampling,
NOTE Confidence: 0.86540186
00:34:54.440 --> 00:34:56.510 so these are going in opposition
NOTE Confidence: 0.86540186
00:34:56.510 --> 00:34:57.545 to each other.
NOTE Confidence: 0.86540186
00:34:57.550 --> 00:34:59.776 Is what I’m saying and then this
relative average standard deviation.

Is this dotted bar line? So what is the main message of this is that these are going in opposite directions right that? The sample that is optimal for the average treatment effect is on the right. The sample that is optimal for moderate effects is on the left, and so there’s tradeoffs involved in these that what’s best involved in these that what’s best for one is not best for the other. But there’s other lessons in here, wow, so the B index is, which is a measure of similarity.
between the sample and population, is actually not that bad for the optimal sample. So these the sample is different from the population. You'd have to do some waiting, but it wouldn't be a tremendous amount of waiting to be able to estimate the average treatment effect. And so one lesson that you could think of it from. This is if you actually if we designed randomized trials to test moderators, we'd actually be in a pretty good space to test moderators. And to estimate the average treatment effect,
It wouldn’t be that far off.

It wouldn’t be.

It wouldn’t be terrible, and that makes sense because we’re covering so much of the population by getting her across a bunch of moderators that we can do so that we can re wait when in a domain in which there’s no act extrapolations, we have positive ITI we can re wait next. Another sort of I think finding here is if we look over at the right hand side. If we do, you know the trade off is. If I do select for the average treatment effect.
I do get a tremendously, you know I can select for the average treatment effect and do pretty well for the average human effect, but not so well for that. For the moderators, and so what’s ideal for average is definitely not deal for the moderate are tests. As I was saying, they actually did a pretty good job in terms of representativeness. You can see that that top dot, but if you look at the bottom
at the other two dots you can see they didn’t do so well for.

Being able to test these moderators. OK, so in case that was not intuitive another way you could look at this is to actually just look at what these samples these features would look like. In the top the top row here are population distributions. Of these five covariates that were sort of identified, and then at the bottom row is actually the study that they had. So what their actual sample looked like.
And then the middle is what AD optimal sample would look like. And then I've overlaid on here. These are values, so giving you a sense if R is greater is 1. It means the sample is like the same standard deviation as in the population. If R is greater than one, it means I've got more heterogeneity in my sample than in my population, which improves my ability to estimate moderate are effects. And So what you see are a few things. One is in that the optimal sample is. It pushes things towards the extremes.
It’s pushing them towards the extremes to get endpoints which we know from basic experimental design, improved abilities. The other nice thing though, is a concern always when you’re doing experimental design like this is that you’re going to get your highly focused on like a linearity assumption that you’re going to your highly focused on like a linearity assumption that you’re going to your. Your ideal sample would have a strong linearity assumption to it, but because you have multiple variables an because not all design runs are possible. In the population,
you end up with these middle points
NOTE Confidence: 0.861584369999999

as well so you don’t end up with only things on both extremes.
NOTE Confidence: 0.861584369999999

You end up with some middle points which allow you to be able to estimate nonlinear relationships as well.
NOTE Confidence: 0.861584369999999

You can see that you would just end up with a lot more variation and so not surprisingly, total students, which, again schools studies, tend to over represent very large schools and large school districts.
NOTE Confidence: 0.861584369999999

You can see this is a place where there would be really a real opportunity for a change that in
the sample this was less than one
an in the optimal sample it would be greater than three.
But you can see this for most of these variables that you could.
You could potentially improve your power and ability to estimate things related to demographics as well.
And in my paper I actually show that because many of these are proportions, you can actually also think about student level moderate yrs because proportions conveniently like the variation in proportions at the individual level as a function of.
the proportion at the aggregate.

And so you can actually kind of work out a way to select your samples so that you can estimate individual affects, not just cluster aggregates for those variables.

OK, and so then the final point. I just want to make is that the other thing that this shows is that there’s real benefit to augmentation so. Maybe? You know, maybe I’m not going to be able to convince people to go switch to selecting their samples based upon extremes. But maybe you can convince people that they could preserve 5 or 10.
You know 10% or 25% of their sample for D optimality. So you choose. In this case it would be like choose 30 of your sites using stratified sampling to represent the population, and then look for like an additional class tenor 11 sites that might be more extreme that allow you to make sure that you can estimate these. These moderate are effects that you’re interested in. And you can see that doing so key file with these little lines you can see that doing so doesn’t have a huge
effect on the average treatment effect, but it does greatly improve your ability to test moderators.

OK, so just to wrap up my take home points today, I suppose would be that the design of randomized trials has big implications for ability to generalize. And I think we, I think what I’ve seen over time is that people who are starting to pay attention to that, and they’re starting to think about how populations you know. What are the populations I would add as a side benefit of this is I’ve watched as people in asking people to scientists to think
about what the population is.

It actually sometimes make some change with the intervention is because you kind of have to realize like is this is this.

If this is the population,

If this is the right intervention?

The second sort of point I would say, that if we want to sort of estimate and test hypothesis and moderators that we would be wise to actually plan to do so and to think about how to have better design sensitivity and statistical power for doing so instead of waiting until the end and then the last point is.
just that this augmentation approach indicates that we don’t have to be perfect at this like that, we could just, you know, use do this for part of our sample. And we would be better off and then I guess I would say maybe my general philosophy in all of this design is that. What I’m trying to do is to get people to think differently and plan differently, and by doing so, even if you don’t succeed 100%, you’re better off than you would have been before, and you’re now able to be in the realm in which you have positive
ITI and heterogeneity, and you're able to actually use statistical methods.

To get better estimators at the end.

Thank you, this is all my contact information and this is the paper that this talk is really about.

I’m happy to answer questions.

I think that’s really nice talk and thank you for being so inspiring.

And maybe let’s open to questions 1st to see if we have any questions from the audience.

If so, please speak up or, you know,
send a chat. Either one is OK.

And if not, I can go first.

'scause I do have a couple of questions.

So, so first of all, I think you know there is a constant tension.

Of course, like you know when we work with really large trials in the healthcare system,

I think there is a tension between how do we better represent the population of interest?

Because we want to get effectiveness information 'cause we're spending millions of dollars.

But also I think there is a concern on you know how to really better engage these large clusters,

large healthcare systems or large clinics,
And so I think.

People end up getting convenience samples because that’s reality.

Even though I do believe that there’s so much more to improve because they’re spending so much money right.

And then in the end, you know they may be answering a different question if they have a very highly selected sample and then people also worry about you know, like you know there are some disparities in their sample selection, so that you’re basically not covering people with maybe more
vulnerable conditions etc in your study,

but you wish to answer questions.

What is population?

So I feel like all of this very,

very relevant, at least to my work.

And so I really appreciate you know this

aspect of how to design styles better.

1 one of the questions I have

is that generally,

you know we may not really know priority

what the effect modifiers are in planning.

The trial that we may have

not enough knowledge amount.

So how does that generally come into

the discussion in the design stage?

Is it the tradition that in educational
studies we have a lot of prime knowledge on what these effect modifiers are or? No, so I think this is actually one of the hardest parts, right? Sort of, if we knew what the why zeros and Y ones were, this is what we would. You know this is that would be optimal, but I could be wrong on what those are, right? And I don’t know. I mean, I think so. There’s sort of what I call the usual suspects in education,
which are like race class and gender,
NOTE Confidence: 0.8577501
which are really more of concerns
NOTE Confidence: 0.8577501
about disparity or about closing
NOTE Confidence: 0.8577501
achievement gaps in various ways.
NOTE Confidence: 0.8577501
And so those in depth and
NOTE Confidence: 0.8577501
urbanicity I would add seems to be
NOTE Confidence: 0.8577501
something that people often like.
NOTE Confidence: 0.8577501
What add into that as characteristics.
NOTE Confidence: 0.8577501
Those are the ones that
NOTE Confidence: 0.8577501
people most often use.
NOTE Confidence: 0.8577501
But the and those are
NOTE Confidence: 0.8577501
available in population data,
NOTE Confidence: 0.8577501
which is the other thing
NOTE Confidence: 0.8577501
that your limit your.
NOTE Confidence: 0.8577501
A real limiter is what is available
NOTE Confidence: 0.8577501
in the population, sure.
What I gather is more likely to be a moderate achievement or something like baseline achievement, right? So if my outcome is achievement then I would think that what the achievement is baseline in any of these places would matter. That’s harder to get an education. I mean that information from places, so there’s been some work trying to equate tests across across states. I guess that they do. Sometimes they use gain scores just to subtract off that baseline achievement, right? They
Note: Confidence: 0.851321227777778

00:46:13.010 -- 00:46:13.910 do. Yeah, exactly,

00:46:13.910 -- 00:46:16.392 but the problem is that like if you

00:46:16.392 -- 00:46:18.646 wanted to use state tests or something,

00:46:18.650 -- 00:46:20.778 there are different tests in every state,

00:46:20.780 -- 00:46:22.694 and so there’s all of these

00:46:22.694 -- 00:46:24.730 equating issues that go in with it.

00:46:24.730 -- 00:46:26.746 My guess is that implementation is another

00:46:26.746 -- 00:46:29.240 one that people often come up with is

00:46:29.240 -- 00:46:30.500 like something with implementation.

00:46:30.500 -- 00:46:32.528 Now this is tricky because implementation

00:46:32.528 -- 00:46:34.178 is coming after assignment and

00:46:34.178 -- 00:46:35.666 so it’s really like a mediator.

00:46:35.670 -- 00:46:37.500 But if you think about often,

00:46:37.500 -- 00:46:38.524 if you think implementation

00:46:38.524 -- 00:46:40.502 may be part of what is leading

00:46:40.502 -- 00:46:42.058 to treatment effect variation,
then you can kind of think well what.

Affects implementation and so

people can sometimes think a little

more carefully about what affects

like Oh well,

it’s probably.

You might try to find various

measures of this for the

implementation that sounds more like a.

It’s sort of a version of multiple

treatments, and it’s a violation of

the suitable condition, probably.
00:47:12.392 --> 00:47:13.480 Yeah, yeah, exactly yeah.
NOTE Confidence: 0.85544264

00:47:13.480 --> 00:47:15.384 So I mean so it it gets.
NOTE Confidence: 0.85544264

00:47:15.390 --> 00:47:17.016 It gets tenuous. Yeah, I don’t.
NOTE Confidence: 0.85544264

00:47:17.020 --> 00:47:18.916 I don’t have this is, you know,
NOTE Confidence: 0.85544264

00:47:18.916 --> 00:47:21.060 this is like I when I first started
NOTE Confidence: 0.85544264

00:47:21.121 --> 00:47:22.729 doing this work I was like,
NOTE Confidence: 0.85544264

00:47:22.730 --> 00:47:24.606 well assuming moderate yrs and assume a
NOTE Confidence: 0.85544264

00:47:24.606 --> 00:47:26.270 population moving on as a statistician.
NOTE Confidence: 0.85544264

00:47:26.270 --> 00:47:27.896 But actually those are the two
NOTE Confidence: 0.85544264

00:47:27.896 --> 00:47:29.301 hardest things when working with
NOTE Confidence: 0.85544264

00:47:29.301 --> 00:47:30.676 people in planning these trials
NOTE Confidence: 0.85544264

00:47:30.676 --> 00:47:32.250 is thinking about what they are.
NOTE Confidence: 0.85544264

00:47:32.250 --> 00:47:33.876 I’ll give you an example though.
NOTE Confidence: 0.85544264

00:47:33.880 --> 00:47:36.645 Uh, like a positive case which was.
NOTE Confidence: 0.85544264

00:47:36.650 --> 00:47:39.086 I was part of designing something called
NOTE Confidence: 0.85544264

00:47:39.086 --> 00:47:41.430 the National Study of learning mindsets,
which is we randomly sampled 100 high schools in the US, and then we randomly assigned to either using a computer based intervention to a growth mindset intervention or something that was not growth mindset that was just sort of control condition and. And in doing that we had the social psychologist I was working with had a lot of questions like we had a lot of hard questions about these.
moderators and they had a lot of theories about what they might be like. So we oversampled like we. Looked at for example, proportion of students that are minorities in the school and then. And when we started we wanted to stratify on that as well as school achievement as well, and so we needed to be able to cross these in a way in order to. In order to D alias these trends and so that they could estimate one without, you know, without estimating with separated
from the other. So a lot of it.

So I mean, in some places people are much more better,
much better theoretically.

Thinking about this, I think some fields are better at
thinking about these mechanisms than
other fields are, but yeah, it’s really hard.

So my other than my
like standard set, you know,
race, class and gender is,
I often ask people to to think about.
Watch what variables might just be
related to other things, right?
That if you could.

If you can think of it as like I ultimately want to test moderators that I don’t really know exactly but I need to get variation in them, and that means probably by getting variation in something else.

The size of your site, you know, is one place where you know an education. You can see that everybody’s in very large sites.

And So what if we increase the variation?
The variation of district size and school size? It seems like it has to increase variation of some other things as well. Agreed, agreed.

Yeah, I think another aspect why I so appreciate like the aspect of effect modifiers is that it really is a way to move forward with information from Co. Buryats and then when we talk on 80 in a randomized study, we often ignore covariates and then just hold that the unadjusted analysis provides unbiased estimates, even though that may come with
00:50:12.888 --> 00:50:14.022 a larger variation.
NOTE Confidence: 0.83546096
00:50:14.030 --> 00:50:16.690 So by really talking about effect modifiers,
NOTE Confidence: 0.83546096
00:50:16.690 --> 00:50:18.580 we somehow incurve those information,
NOTE Confidence: 0.83546096
00:50:18.580 --> 00:50:20.854 but perhaps even in the estimation
NOTE Confidence: 0.83546096
00:50:20.854 --> 00:50:22.370 of the average affect,
NOTE Confidence: 0.83546096
00:50:22.370 --> 00:50:23.890 which can increase precision.
NOTE Confidence: 0.83546096
00:50:23.890 --> 00:50:24.650 So yeah.
NOTE Confidence: 0.85203755
00:50:26.800 --> 00:50:27.070 Yeah.
NOTE Confidence: 0.79443485
00:50:30.360 --> 00:50:32.112 Yeah I haven’t questioned,
NOTE Confidence: 0.79443485
00:50:32.112 --> 00:50:34.740 so I actually have two questions,
NOTE Confidence: 0.79443485
00:50:34.740 --> 00:50:36.930 so seems you’re you’re interested
NOTE Confidence: 0.79443485
00:50:36.930 --> 00:50:38.678 in both individual level
NOTE Confidence: 0.79443485
00:50:38.680 --> 00:50:41.310 an cluster level moderators right? When
NOTE Confidence: 0.79443485
00:50:41.310 --> 00:50:43.500 you have cluster level moderators,
NOTE Confidence: 0.79443485
00:50:43.500 --> 00:50:45.690 how does that work with
NOTE Confidence: 0.79443485
00:50:45.690 --> 00:50:47.002 the augmentation design?
'cause you mentioned that in the orientation design, you might want to pick like. An kind of like choose them samples from those. But how do you choose those 3%? You choose those third percent with respect to the cluster level modelers. You could do it with respect to either. You can do it with respect there because it depends the way you enter them into the model so. You work out so you can work out that if I'm interested in the individual level.
Moderate are that that what I need to do is I need the interaction between like X an 1 -- X. That’s what I included here as the covariates. I’m ’cause I want to increase the variation within sites right? And so you could do it either way, because what it’s doing, what the augmentation approach does? Is it assess is how much variation you have in those 30 sites already. And then it looks for possible design runs, meaning other samples. Other places that would greatly improve that.
And it just doesn’t algorithmically, which is nice.

The that I would say I should add an extra benefit of this is concerned with all of this sample recruitment is that there’s non response. You’re never going to get, you know it’s not like I can just say like. Here’s your. Here’s your like 40 sites go ask them and they’re going to say yes, but with the augmentation approach if somebody says no you can like throw that out and then go look for it like what’s the next best alternative.
so you can keep kind of iterating.

So, I think the attributes are all cluster level.

Information right summary statistics

That’s in the slides but

That’s in the in the slides but

I didn’t include in here though

is like you could it’s but it’s

in the paper is you could also do

this with individual level only.

For proportions mean because just because

the proportion works out that you can get.

You can think about this with

the same statistics you would
get at the cluster level.

You can’t get the variation you can with a normal like a continuous variable.

I can’t get the standard deviation.

I don’t have the standard deviation.

Insights I can’t do that, right, right?

Also, the other question is about so it seems like all these designs are under the assumption that you’re interested in all the moderators, interested in all the moderators, like equally like meaning that you’re not like you don’t have like primary moderators that you’re interested in.
estimating the moderate effect on and then you have a couple of them that. I mean, if you can. So I mean, what’s great?

I think about this like this area is that it’s been so richly developed in this other sort of design runway is that you can actually add weights. So you can say like I’m more like or more interested in this variable than that variable, and it will focus. You know it will focus on one variable over the other. Because you Can you imagine like that D matrix. The determinant of S.
You could just add weights into that. So if you add weights into that then you can start looking at the determinant of that weighted version. Right, so you would add weights in that matrix and optimize that. Yeah exactly, if you add weight so that some of the Kobe rates are getting more weight than others. So I guess just maybe more precisely, I think the D optimality criteria. Shouldn’t that be the X transpose V universe in general? Just because you’re working with
clustered randomized studies so that the outcome correlation is somehow included in that variance? Is that what the algorithm is trying to get in general for? Yeah, Inverse, yeah, it’s the X prime X inverse, which is the covariance. Yeah, but but really not so it you don’t. You don’t need to have the variance matrix of the outcome. Exactly, you don’t need to have the outcome, it’s all about the inputs, right? But that’s, which is why you can do it in advance, right? So it’s all about the Android just nicely.
You can leverage population data that you have totally. And again, I assume in all of this there's measurement error and that you know you can just sort of assume that like you're not going to get it exactly right, but my baseline comparison is always what are we doing now versus what could we be doing an like frankly anything. Any you know it looks to me like we have fairly homogeneous samples and any effort we can make to increase heterogeneity is an improvement. So, well, I think we're about the hour.
but let’s see if we have any final questions from the audience. Alrighty, if not, I think you know I’m, I’m sure if you have any questions that petition will we have to answer them offline by email? So thanks so much. Again, bath. It’s really nice to have you and thanks to everybody for attending or see all of you. Hopefully after the break so have a great holiday. See you later. Totally not master connect. Alright, thanks again. Talk to you later. Bye take care.