So welcome everyone, it is my great pleasure to introduce our seminar speaker today, Doctor Elizabeth Tipton. She’s an associate professor and a faculty fellow in the Institute for Policy Research at Northwestern University. Unducted sentence research focuses on the design and analysis of randomized experiments with a focus on issues for 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. 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? How do you usually set this up? Do people what’s the norm? Do you guys usually up 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. OK so I just want to set out background for what I’m talking about today, which is I’m talking about randomized
00:01:34.310 --> 00:01:36.417 I realized that in a Biostatistics Department, you guys.
00:01:36.417 --> 00:01:38.740 The idea that randomized trials are common is probably almost absurdly basic for the world that you operate in,
00:01:38.740 --> 00:01:40.900 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.
00:01:40.900 --> 00:01:43.046 Randomized trials have only become common really,
00:01:43.046 --> 00:01:45.293 I’d say the last 20 years.
00:01:45.300 --> 00:01:47.452 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, sort of somewhat small samples of clusters 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 aniff 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 and 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 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
00:07:21.338 --> 00:07:23.180 sampling language for the same thing.

NOTE Confidence: 0.8677029

00:07:23.180 --> 00:07:24.720 That’s right, yeah, and so.

NOTE Confidence: 0.8677029

00:07:24.720 --> 00:07:26.869 And that’s the lack of positive ITI

NOTE Confidence: 0.8677029

00:07:26.869 --> 00:07:28.535 often arises because people aren’t

NOTE Confidence: 0.8677029

00:07:28.535 --> 00:07:30.605 thinking about what the population is

NOTE Confidence: 0.8677029

00:07:30.605 --> 00:07:32.638 in advanced and so it’s very tricky

NOTE Confidence: 0.8677029

00:07:32.638 --> 00:07:34.550 for them after the fact to generalize,

NOTE Confidence: 0.8677029

00:07:34.550 --> 00:07:36.410 because it turns out that maybe

NOTE Confidence: 0.8677029

00:07:36.410 --> 00:07:38.249 this what I as analyst him now

NOTE Confidence: 0.8677029

00:07:38.249 --> 00:07:40.284 trying to think of as the population

NOTE Confidence: 0.8677029

00:07:40.284 --> 00:07:41.908 isn’t exactly the population,

NOTE Confidence: 0.8677029

00:07:41.908 --> 00:07:43.905 but it’s very hard for people to

NOTE Confidence: 0.8677029

00:07:43.905 --> 00:07:45.589 articulate what the population is,

NOTE Confidence: 0.8677029

00:07:45.590 --> 00:07:48.810 and so spent a lot of time just trying to.

NOTE Confidence: 0.8677029

00:07:48.810 --> 00:07:50.826 You’re out what the population actually is,

NOTE Confidence: 0.8677029
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 what these variables are, because studies have not been designed to estimate or test hypothesis about moderation, and so instead we have to sort of think
through what we think might matter.

Using, you know, not a great source of knowledge here, but the idea is that you sort of take all of this information and then you use this to create to use sampling methods to actually design recruitment procedures like using stratified sampling.

Figuring out if you should you know within Strata, using balanced sampling or random sampling and thinking about sort of ways in which you can increase the
coverage so you can have positive
NOTE Confidence: 0.881084
ITI for the whole target population,
NOTE Confidence: 0.881084
so that when you do you know.
NOTE Confidence: 0.881084
When you do need to make adjustments
NOTE Confidence: 0.881084
at the end of your trial using
NOTE Confidence: 0.881084
these statistical methods,
NOTE Confidence: 0.881084
they are in a realm in which
NOTE Confidence: 0.881084
they can perform well.
NOTE Confidence: 0.8562651
This this sort of lad me to thinking about
NOTE Confidence: 0.8562651
tools for general for generalization,
NOTE Confidence: 0.8562651
and so I just want to highlight this
NOTE Confidence: 0.8562651
because I think this is a good strategy
NOTE Confidence: 0.8562651
for methods people to think about.
NOTE Confidence: 0.8562651
So I I thought will nobody is going to do
NOTE Confidence: 0.8562651
what I’m telling them to do if I don’t
NOTE Confidence: 0.8562651
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 orsan. 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
00:11:30.730 --> 00:11:32.130 to recruit with the.

Great, I've had this going since 2015 and it was very slow going for awhile.

This is sort of. I just realized this 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
thinking about this.

Um?

So everything I’ve talked about is sort of averaging over hedge and 80 when we talk about analyze.

Estimate an average treatment effect for a population, assuming that there’s variation of effects and we’re averaging over those. But to average over those requires that we know something about how treatment affects very and very often, and I would say this is the in general we don’t, and the reason that we don’t have a
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 heterogeneity. So I’m just going to start with like a little bit of a background here.
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 years 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 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 cluster randomized randomized block individual randomizer, 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, so those are different. And then it’s a function of $N$, 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 is $\sigma_X K^2$?
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,
NOTE Confidence: 0.8265885

moderate yrs and individual
NOTE Confidence: 0.8265885

level moderate yrs and.
NOTE Confidence: 0.8265885

I'm taking basically what they've got,
NOTE Confidence: 0.8265885

but re tweaking part of it.
NOTE Confidence: 0.851218459999999

Because I'm factoring out this Sigma
NOTE Confidence: 0.851218459999999

squared and noting that you can
NOTE Confidence: 0.851218459999999

actually pull out this thing called
NOTE Confidence: 0.851218459999999

RXK at the front and the RXK is
NOTE Confidence: 0.851218459999999

this ratio of the standard deviation
NOTE Confidence: 0.851218459999999

of the covariate in the sample
NOTE Confidence: 0.851218459999999

compared to the standard deviation
NOTE Confidence: 0.851218459999999

of the covariate in the population.
NOTE Confidence: 0.851218459999999

And so what you can see here is by
NOTE Confidence: 0.851218459999999

doing that you by rewriting it.
NOTE Confidence: 0.851218459999999

This way you can see that our XK is
having just as much of an effect on statistical power as things like the square root of N or the square root of P.

These other parameters that most power analysis has spent has focused on, and that’s true. 

You know, in any of these designs. Love seeing it in any of these designs.

RX shows up. OK, so if RX is something that matters for power, a question will be well.

What are people doing in practice right now right? So maybe maybe people are choosing 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 of let’s say elementary schools. I’m looking at the ratio of this 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,
don't?

OK, what you can see here is that the bar at the bottom.

Can you see my cursor?

Can't tell if you guys can see my cursor.

No,

you can't see my cursor.

OK,

and so you can see that most 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

my sample size of 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 heterogeneity matters, that is something that is not actually happening in practice.

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

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 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 do this, 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 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 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're going to have to anticipate this in advance and think carefully about sort of compromises, 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?

OK, so this leads to a principle found in response surface models called D optimality and so AD optimal design. This is work from the 40s and 60 Forties,
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. So 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 30 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.
runs that we’ve chosen to use.

OK, so I’m just going to go through an example to talk about this.

Don’t have a ton more slides. I should say so success.

OK, so here’s an example. The success for all evaluation was an elementary school reading program evaluated between 2001 and 2003. The reason I like to use this example. Is that it’s old enough that strangely, they actually published in their paper a list of schools they actually named the schools in their study and characteristics of them.
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 be 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 of the students. 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
proportion at free and reduced lunch.
And they also talk about Urbanicity because they tried to make sure they
had some urban schools in rural
So I should say in previous work of mine I’ve used this as an example
and then this study actually ends up being a fairly representative
sample of the population,
which is interesting.
Is it because they had no real way
of they weren’t doing it totally in
a way that allowed them to compare
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.

So what I did 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

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.
to measures of undercoverage and how and

NOTE Confidence: 0.86540186

the performance of reweighting methods.

NOTE Confidence: 0.86540186

And then the mean are meaning the ratio

NOTE Confidence: 0.86540186

between the the ratio between the

NOTE Confidence: 0.86540186

standard deviation in the sample and

NOTE Confidence: 0.86540186

population across these five covariates.

NOTE Confidence: 0.86540186

OK, so this is what we get out of this,

NOTE Confidence: 0.86540186

and so I just want to talk through this and

NOTE Confidence: 0.86540186

I’m happy to answer questions if there’s.

NOTE Confidence: 0.86540186

I know there’s a lot going on here.

NOTE Confidence: 0.86540186

Really wish I could figure out how to do a.

NOTE Confidence: 0.86540186

Pointer.

NOTE Confidence: 0.86540186

I don’t think I can point out that way.

NOTE Confidence: 0.86540186

OK, so OK.

NOTE Confidence: 0.86540186

So what I have going on here is the number

NOTE Confidence: 0.86540186

do 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 so 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 are effects is on the left,

and so there’s tradeoffs 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 these the sample is different from the population. You’d have to do some re waiting, but it wouldn’t be a tremendous amount of re 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 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, what's ideal for average is definitely not deal for the moderate are tests. So and then the third thing would be if you look at the actual study. 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
00:36:56.390 --> 00:36:58.594 at the other two dots you can
00:36:58.594 --> 00:37:00.800 see they didn’t do so well for.
00:37:00.800 --> 00:37:03.229 Being able to test these these moderators.
00:37:06.550 --> 00:37:09.142 OK, so in case that was not intuitive
00:37:09.142 --> 00:37:11.420 another way you could look at this
00:37:11.420 --> 00:37:13.830 is to actually just look at what
00:37:13.830 --> 00:37:15.785 these samples these these features
00:37:15.785 --> 00:37:20.210 So in the top the top row here
00:37:20.210 --> 00:37:21.818 are population distributions.
00:37:21.820 --> 00:37:23.550 Of these five covariates that
00:37:23.550 --> 00:37:24.934 were sort of identified,
00:37:24.940 --> 00:37:27.341 and then at the bottom row is
00:37:27.341 --> 00:37:29.450 actually the study that they had.
00:37:29.450 --> 00:37:32.537 So what their actual sample looked like.
And then the middle is what AD
NOTE Confidence: 0.861584369999999
00:37:34.808 --> 00:37:36.479 optimal sample would look like.
NOTE Confidence: 0.861584369999999
00:37:36.480 --> 00:37:38.436 And then I've overlaid on here.
NOTE Confidence: 0.861584369999999
00:37:38.440 --> 00:37:39.433 These are values,
NOTE Confidence: 0.861584369999999
00:37:39.433 --> 00:37:43.037 so giving you a sense if $R$ is greater is 1.
NOTE Confidence: 0.861584369999999
00:37:43.040 --> 00:37:45.656 It means the sample is like the same
NOTE Confidence: 0.861584369999999
00:37:45.656 --> 00:37:47.630 standard deviation as in the population.
NOTE Confidence: 0.861584369999999
00:37:47.630 --> 00:37:49.598 If $R$ is greater than one,
NOTE Confidence: 0.861584369999999
00:37:49.600 --> 00:37:51.514 it means I've got more heterogeneity
NOTE Confidence: 0.861584369999999
00:37:51.514 --> 00:37:53.859 in my sample than in my population,
NOTE Confidence: 0.861584369999999
00:37:53.860 --> 00:37:55.495 which improves my ability to
NOTE Confidence: 0.861584369999999
00:37:55.495 --> 00:37:56.803 estimate moderate are effects.
NOTE Confidence: 0.861584369999999
00:37:56.810 --> 00:37:59.753 And so what you see are a few things.
NOTE Confidence: 0.861584369999999
00:37:59.760 --> 00:38:02.488 One is in that the optimal sample is.
NOTE Confidence: 0.861584369999999
00:38:02.490 --> 00:38:04.686 It pushes things towards the extremes,
NOTE Confidence: 0.861584369999999
00:38:04.690 --> 00:38:05.048 right?
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. Your ideal sample would have a strong linearity assumption to it, but because you have multiple variables and because not all design runs are possible. In the population,
you end up with these middle points
NOTE Confidence: 0.861584369999999
so you don’t end up with
NOTE Confidence: 0.861584369999999
only things on both extremes.
NOTE Confidence: 0.861584369999999
You end up with some middle points
NOTE Confidence: 0.861584369999999
which allow you to be able to estimate
NOTE Confidence: 0.861584369999999
nonlinear relationships as well.
NOTE Confidence: 0.861584369999999
Me and a Third Point with me.
NOTE Confidence: 0.861584369999999
You can see that you would just end
NOTE Confidence: 0.861584369999999
up with a lot more variation and so
NOTE Confidence: 0.861584369999999
not surprisingly, total students,
NOTE Confidence: 0.861584369999999
which, again schools studies,
NOTE Confidence: 0.861584369999999
tend to over represent very large
NOTE Confidence: 0.861584369999999
schools and large school districts.
NOTE Confidence: 0.861584369999999
You can see this is a place where
NOTE Confidence: 0.861584369999999
there would be really a real
NOTE Confidence: 0.861584369999999
opportunity for a change that in
NOTE Confidence: 0.861584369999999
00:39:05.447 --> 00:39:07.225 the sample this was less than one
NOTE Confidence: 0.861584369999999
00:39:07.225 --> 00:39:09.330 an in the in the optimal sample
NOTE Confidence: 0.861584369999999
00:39:09.330 --> 00:39:11.340 it would be greater than three.
NOTE Confidence: 0.861584369999999
00:39:11.340 --> 00:39:13.908 But you can see this for most of
NOTE Confidence: 0.861584369999999
00:39:13.908 --> 00:39:15.630 these variables that you could.
NOTE Confidence: 0.861584369999999
00:39:15.630 --> 00:39:17.250 You could potentially improve your
NOTE Confidence: 0.861584369999999
00:39:17.250 --> 00:39:19.260 power and ability to estimate things
NOTE Confidence: 0.861584369999999
00:39:19.260 --> 00:39:20.910 related to demographics as well.
NOTE Confidence: 0.861584369999999
00:39:20.910 --> 00:39:23.318 And in my paper I actually show that
NOTE Confidence: 0.861584369999999
00:39:23.318 --> 00:39:25.530 because many of these are proportions,
NOTE Confidence: 0.861584369999999
00:39:25.530 --> 00:39:27.402 you can actually also think about
NOTE Confidence: 0.861584369999999
00:39:27.402 --> 00:39:29.088 student level moderate yrs because
NOTE Confidence: 0.861584369999999
00:39:29.088 --> 00:39:30.608 proportions conveniently like the
NOTE Confidence: 0.861584369999999
00:39:30.608 --> 00:39:32.508 variation in proportions at the
NOTE Confidence: 0.861584369999999
00:39:32.508 --> 00:39:34.280 individual level as a function of
NOTE Confidence: 0.861584369999999
the proportion at the aggregate.

And so you can actually kind of workout 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 10 sites that might be more extreme that allow you to make sure that you can estimate these. These moderate are effects. 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 that 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,
is 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 and maybe let’s open to questions and maybe let’s open to questions and maybe let’s open to questions.
00:42:58.146 --> 00:43:00.648 send a chat. Either one is OK.
NOTE Confidence: 0.9289141
00:43:07.090 --> 00:43:08.966 And if not, I can go first.
NOTE Confidence: 0.9289141
00:43:08.970 --> 00:43:12.514 'cause I do have a couple of questions.
NOTE Confidence: 0.9289141
00:43:12.520 --> 00:43:15.330 So, so first of all, I think you
NOTE Confidence: 0.9289141
00:43:15.330 --> 00:43:17.430 know there is a constant tension.
NOTE Confidence: 0.9289141
00:43:17.430 --> 00:43:20.238 Of course, like you know when we work with
NOTE Confidence: 0.9289141
00:43:20.238 --> 00:43:23.047 really large trials in the healthcare system,
NOTE Confidence: 0.9289141
00:43:23.050 --> 00:43:25.820 I think there is a tension between how do we
NOTE Confidence: 0.9289141
00:43:25.892 --> 00:43:28.670 better represent the population of interest?
NOTE Confidence: 0.9289141
00:43:28.670 --> 00:43:30.770 Because we want to get effectiveness
NOTE Confidence: 0.9289141
00:43:30.770 --> 00:43:31.820 information 'cause we're
NOTE Confidence: 0.9289141
00:43:31.820 --> 00:43:33.230 spending millions of dollars.
NOTE Confidence: 0.9289141
00:43:33.230 --> 00:43:35.534 But also I think there is a concern
NOTE Confidence: 0.9289141
00:43:35.534 --> 00:43:38.247 on you know how to really better
NOTE Confidence: 0.9289141
00:43:38.247 --> 00:43:39.895 engage these large clusters,
NOTE Confidence: 0.9289141
00:43:39.900 --> 00:43:42.000 large healthcare systems or large clinics,
00:43:42.000 --> 00:43:44.010 etc. And so I think.
00:43:44.010 --> 00:43:45.625 People end up getting convenience samples because that’s reality.
00:43:45.625 --> 00:43:48.922 Even though I do believe that there’s so much more to improve because they’re spending so much money right.
00:43:50.910 --> 00:43:53.046 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, 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? Like I just laid out. 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, which are really more of concerns about disparity or about closing achievement gaps in various ways. And so those in depth and urbanicity I would add seems to be something that people often like. What add into that as characteristics. Those are the ones that people most often use. But the and those are available in population data, which is the other thing that your limit your. A real limiter is what is available in the population, sure.
What I gather is more likely to be a moderate achievement, right? 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
do. Yeah, exactly,

but the problem is that like if you

wanted to use state tests or something,

there are different tests in every state,

and so there’s all of these

My guess is that implementation is another

one that people often come up with is

like something with implementation.

Now this is tricky because implementation

is coming after assignment and

so it’s really like a mediator.

But if you think about often,

if you think implementation

may be part of what is leading

to treatment effect variation,
00:46:42.060 --> 00:46:44.588 then you can kind of think well what.

00:46:44.590 --> 00:46:45.922 Affects implementation and so

00:46:45.922 --> 00:46:47.920 people can sometimes think a little

00:46:47.978 --> 00:46:49.743 more carefully about what affects

00:46:49.743 --> 00:46:51.155 implementation like Oh well,

00:46:51.160 --> 00:46:51.902 it’s probably.

00:46:51.902 --> 00:46:54.499 You know, it’s probably easier to implement

00:46:54.499 --> 00:46:56.696 this in schools that are like this.

00:46:56.700 --> 00:47:00.536 Then schools that are not like that.

00:47:00.540 --> 00:47:02.658 You might try to find various

00:47:02.658 --> 00:47:04.360 measures of this for the

00:47:04.360 --> 00:47:07.768 implementation that sounds more like a.

00:47:07.770 --> 00:47:09.674 It’s sort of a version of multiple

00:47:09.674 --> 00:47:11.305 treatments, and it’s a violation of

00:47:11.305 --> 00:47:12.392 the suitable condition, probably.
Yeah, yeah, exactly yeah.
So it gets tenuous. Yeah, I don’t.
I don’t have this is, you know,
this is like I when I first started
assuming moderate yrs and assume a
population moving on as a statistician.
But actually those are the two
hardest things when working with
people in planning these trials
is thinking about what they are.
I’ll give you an example though.
Uh, like a positive case which was.
I was part of designing something called
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 at a measure of sort of 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
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 my other 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 what they are, but I need to get variation in them, and that means probably by getting variation in something else. I’m going to get variation in those as well, I think. 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 has to increase variation of some other things as well. 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
a larger variation.

So by really talking about effect modifiers, we somehow incurve those information, but perhaps even in the estimation of the average affect, which can increase precision.

Yeah.

Yeah I haven’t questioned, so I actually have two questions, so seems you’re you’re interested in both individual level and cluster level moderators right? When you have cluster level moderators, how does that work with 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. 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 I actually need to include as a covariate the interaction between like X an 1 -- X. That’s what I included here. 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 it 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, in our current application, I think the attributes are all cluster level. Information right summary statistics

Yeah yeah, well that’s what I have right here. That’s in the in the slides but it’s you could also do this with individual level only.

You can think about this with the same statistics you would.
00:52:51.675 --> 00:52:53.145 get at the cluster level.

00:52:53.150 --> 00:52:55.302 You can’t get the variation you can with

00:52:55.302 --> 00:52:57.387 a normal like a continuous variable.

00:52:57.390 --> 00:52:58.642 I can’t get the.

00:52:58.642 --> 00:53:00.520 I don’t have the standard deviation.

00:53:00.520 --> 00:53:02.510 Insights I can’t do that,

00:53:02.510 --> 00:53:04.850 right, right?

00:53:04.850 --> 00:53:06.150 Also, the other question

00:53:06.150 --> 00:53:07.778 is about so it seems

00:53:07.780 --> 00:53:09.400 like all these designs are

00:53:09.400 --> 00:53:11.028 under the assumption that you’re

00:53:11.030 --> 00:53:12.650 interested in all the moderators,

00:53:12.650 --> 00:53:14.600 like equally like meaning that you’re

00:53:14.600 --> 00:53:16.876 not like you don’t have like primary

00:53:16.876 --> 00:53:18.505 moderators that you’re interested in

NOTE Confidence: 0.8673171

100
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?
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, 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 that like 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 that 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. Totally not master connect.
Talk to you later. Bye take care.