[SQUEAKING]

[RUSTLING]

[CLICKING]

ESTHER DUFLO: Jumping in where we left off, so we basically discussed the Wald estimate, which is a ratio of the difference in mean for the outcome for the group that is the one for the instrument minus the group that has 0 divided by the same thing for the intervening variable-- so the ratio of the first-- the reduced form divided by the first stage.

Now, we could have done that in a regression framework. If you remember, when we run a regression, when we run a variable on a dummy variable, simply, we know that my 1 here is going to be numerically the difference in mean anywhere, and same thing for the gamma 1. So we could have run these two regressions taken by 1 and gamma 1, taken the ratio, and it would have been the Wald estimate.

Another way we could have done that-- and I'm sort of building up to what if, for example, the first stage is not 1, 0, then it's not going to be easy to do the Wald-- not the first stage, the Zi instrument is not 1, 0, then we can't easily do the Wald. So we have to build up to that.

So what we could have done in this case is to say, well, run the first stage, take the predicted value for alpha, for Ai, and then run a regression of Yi on this predicted value.

And, again, the point estimate of beta here would have been the Wald estimate. So let me prove that to you. The point estimate of beta would have been the ratio of the covariance between y and A hat divided by the variance of A hat. That's by definition of what the OLS is.

I'm replacing A hat by its value. A hat is simply pi 0 plus pi 1 Zi. We divide by pi 0 plus pi 1 Zi. Then I can take-- by the properties of variance and covariance, I can take some terms out. The covariance of Y and pi 0 is, of course, 0 because pi 0 is a constant.

The covariance of pi Yi and pi 1 Zi is pi 1 times the covariance of Yi Zi. And then here we have just pi 1 squared variance of Zi. This term here is simply the OLS coefficient of Zi in the Yi regression. So that's gamma 1.

And we are left with an extra pi 1 that we use here. So this is the ratio of gamma 1 over pi 1, which is the ratio of the OLS estimate, which is the Wald estimates. OK?

So, basically, instead of running the Wald estimates, we could do what is literally two state Least Squares. First stage, you run Ai on the instrument. Second stage, you take the predicted value, and you stick it in there.

And the nice thing about it is that you could do that even if you don't have-- even if the Zi is not 1, 0 anymore. You can do that for any type of Zi. And that's called two state Least Squares.

So more generally, imagine that your model is the one we started with at the beginning of previous lecture, where Yi is beta 0 plus beta 1 X1, which is the variable that you're interested in and concerned about-- that's the one that you want to instrument for-- and then some control variable X2 that engender region effect, whatever, that you just want to put in the regression, but you don't need to instrument.

So they are control variables. Sometimes, those control variables are needed because the instrument is only valid when these control variables are included. Now you look for an instrument for X.

If X has more than-- if X is a matrix as opposed to being just 1-- so if you have two things that you want to instrument with at the same time, you need at least as many instruments as you have Xs. Or you could also have more, but at least the same.

Remember, our instruments, they need to be good. They need to be, of course, correlated with X1. They need to be randomly assigned, or as good as randomly assigned, so that we can have a causal effect of the instrument on the Y. And they also need to not have a direct effect on the Y So that all their effect on the Y can be confidently attributed to the X.

Let's say we found those instruments. Then denote Z the matrix of the instruments plus the X2. That's very important, that the metrics of instrument include both the variables Z that are instrumenting for the X1 and the stuff that's going to instrument for itself that we are just controlling for.

In the old days, the Stata-- if you're using Stata, the new Stata syntax for IV regression, in my view, is horrible because you forget that because it's telling you you have your Xs instrumented for and some Zs that are instruments for the X1's, and then you have the other stuff that's not-- this is silly. In term of matrix algebra that's underlying this regression, it makes no sense.

R seems to be like the old Stata, doing it properly. But so the matrix of instrument is the stuff that is instrumenting for the X1 and all of the stuff that instruments for itself. Just remember that.

And then, intuitively, what would we do? Well, following what we have done in the case of just one variable, we run X1 on the matrix of instruments, that is, all of the instruments plus all the X2's. And in the second stage, we take the predicted value for X1, and then we put in the X2 as well.

So in practice, if you do that, the Wald estimate will be correct. But the standard error on all the tests will be wrong because if you do that mechanically, you forget in the estimation that you've estimated the first stage, and therefore, there is additional error that comes from that. So you need to take it into account.

So in practice, that's not what you're going to do. In practice, you're going to tell your software, Mr. R, Mr. Stata, that, hey, this is my model, and this is my matrix of instruments, which is going to include whatever is instrumenting for X1 and the X2's, OK?

And then what is R going to do? Well, if you have the same number of instrument as X-- so, for example, if you have one endogenous variable that is instrumented with one instrument, and the Z matrix and the X matrix are the same size, then the 2SLS formula, which is actually done in one step but is intuitively this two-step formula, is Z prime X minus 1 Z prime i.

So what you do is you replace X prime X of the OLS formula with the Z. What does it do? It takes the Xs. It projects it onto the vectors of Zs. And it takes only the part of the Xs that's explained by the Z to instrument with the Y. So that's the formula for 2SLS.

And the variance is sigma squared Z prime X minus 1 Z prime Z Z prime x minus 1. So this guy is a little bit bigger than an OLS case. In the OLS case, we don't have these two, because we have X prime X minus 1, X prime X, X prime X minus 1, so these two go away.

So that's not a very scientific intuition, but that's useful. The larger these matrices are in the middle, the bigger the standard error. So basically, the fact that these things refuse to collapse is basically accounting for the fact that the first stage is estimated, and your standard error are going to be a bit larger than if you pretended that the X hat that we're putting in the regression is known. So this is what's-- this is what's basically reflecting that.

So this is for the case-- this is the formula for the case where you have the same number of instruments as you have variables in the model-- so, for example, one endogenous X and five control and one instrument and five control. If you have more instruments than variable-- for example, it could happen that have two good instrument for-- so you have more Zs than X-- the formula is a little bit more complicated. It's a little bit longer.

No need for you to learn it, or no need for me to put it there, but it's really the same idea. You're taking the X. You're projecting onto the space of Zs. And then you're using this projection as your regressor-- two state Least Squares done in one step. So that's the formula.

Let's see how R does it. So you run a regression. This is for the Ghana example that we saw before. So we are loading the data, et cetera. And then we're running a regression. This is the command. So we say total score is the dependent variables, and we want whether you complete secondary school plus a bunch of region-- for region-fixed effect. So region.f is my fixed effect.

And then these are my instruments. Treatment is instrument for its complete, and then the region-fixed effect. So R does it properly. The syntax of R is correct. You have all your Xs here and all the Zs here, the entire Z matrix here. OK?

And this is what it's going to report, something very similar to the OLS, something very similar to the OLS table. It's going to tell you what was estimated in the first place. Then you get your coefficient for completed secondary school, which is 0.64. So remember, when we did by hand, this is about what we ended up as well.

And then we have the region dummies that are here instrumented as well. We could add more control if we wanted. And here we would expect the control to make no difference because we have a good instrument. We have a randomly assigned instrument. And, in fact, they don't make-- they don't seem to make any-- much difference at all. The coefficient is still 0.63 with a standard error of 0.23, T statistic of 2.7, and the p-value over here very, very low.

So it's showing very large impact of going to secondary school-- I mean, very significant and quite large impact of going to secondary school on scores at the cognitive exam, which is reassuring. If we didn't have that, that would be a bit worrying. That's our formula.

So that's kind of-- I'm done with formula for today, in fact, maybe forever, because on Monday, we'll do visualization. I don't think it involves all that many formulas. But I'm not done with thinking, because today we are going to think about designing experiments on the notion of what people think about when designing experiments.

It might seem a bit-- it is actually not specialized to think about experiments because I think if you're good at-- if you can think about experiments, you can also think about non-- about good nonexperimental methods. So what do we think? What's the experimental design? What are the questions we need to answer when we're thinking about how to design randomized controlled trials?

So first, we need to think about what is being randomized. And that means what are the interventions or interventions that we are looking at. So, again, we need to think about who is being randomized. That could be a what to a who. It could be a school, but I want to have different questions.

So what's the level of randomization? Do we randomize at the individual level, at the school level, at the village level, at the cell level within the body, or whatever, if we could do that? And then the sample of which we randomized-- for example, in the Ghana example, we randomized over people who were eligible for the college scholarship. All of these questions have-- all of these decisions have implications for the interpretation of the results, as we discussed at length.

So then how is randomization introduced? Are you going to do it with your computer? Are you going to do it in the field? Are you going to take balls from an urn, et cetera?

And what are the type of modification you will do to the design? For example, are you going to stratify by creating groups of relatively identical individuals and randomized schools or whatever, and randomize within that or not. And then the last question we need to answer is, how many units will be randomized? And these questions we already discussed when we spoke about testing.

You're going to want to have your experiment to have enough power to answer the questions of interest. And at the same time, you might not want to spend zillions of dollars. So that's going to be how many questions, which I'm not going to answer at all today since we already discussed when we talked about power calculation. So this was already a design question.

So these are the type of issues that-- the type of things that you want to think about in advance. Of course, as you do the experiment, there are many other things that you need to be thinking about as things go along. But these are the type of things you need to think before even going to the field.

And what are we trying to achieve when designing experiments? Like, what are the things that we want to do? Well, first of all, we want to be able to do the project.

So one of the big progress-- one of the big progress that has been made in the last 10, 15 years, I would say, is finding ways to introduce randomization by stealth. So, of course, one of the big progress is A/B testing, which means all of us are constantly subject to experiments that we have no idea we are subject to.

So randomization is very easy in the context of a website because people-- there is a ton of traffic, and you can randomize the pages and what people see and even the prices and stuff like that. But in other settings, until about 15 years ago, in a lot of cases, people would tell you, that's very nice, but I'm not going to randomize. That's not possible in my business, or in my policy business.

And a lot of the progress that have been made is to try and introduce some element of randomization. And the nice thing from the point of view of this class is that often the randomization is not perfect in the sense that it doesn't give you the cleanest experiment. But it gives you some handle with some variation that has to be combined with statistical techniques, in particular, IV, to answer the question you really want to answer.

So that's one of the design of-- one of the goal of designing experiments is making the experiment possible anyways in a context where it might not otherwise have been obvious. And the second, of course, is making sure that it's-- actually the first, maybe the most important one, surely is to make sure that the experiment you're running answer the questions that you're interested in.

It would be very bad to start with an experiment and then think about what question it might possibly answer. So typically, you start with a question, and you design your experiment to answer your question. So what I want to do in the rest of today is give you some examples of both of these things, of thinking about the experimental design.

To answer both of these, I'll start with introducing randomization where it may not otherwise be obvious. And then I'll go over two examples of designs to answer that were done to answer specific questions of interest where it is the experimental design that made this possible as opposed to good data collection or stuff like that.

So the easiest way to randomize, of course, when you can is simply to randomize, to do the simple completely randomized experiment. Take a population of eligible people. That's your sample frame. Your unit of randomization could be people, household, et cetera.

And then use a software to do in your office randomly assign one group to treatment, one group to control, or maybe you have several treatment groups. So that's the easiest thing. If you have this luxury that you can just simply randomize-- and it's often the case, actually, that it's possible-- then you can-- there are usually two questions that you need to answer at the design stage.

One of them is, do you want to stratify, and one of them is, do you want to cluster? Now, almost the opposite, going in opposite direction. So what is stratifying? Stratifying is starting by creating strata.

Think of strata as buckets. Buckets of people that are schools or units or whatever that are going to be treated separately. So why would we need to stratify? A main reason is to randomize within people who are similar.

So, for example, in the Ghana experiment that we saw, we stratified by gender, and we stratified by region, which is why all of the regression had the strata dummies. Sometimes, we stratify because we also want, for example, to give different randomization chances in different strata. I'll talk to that in a moment.

And that's fine, then. It means the probability to be assigned to the treatment depends on your potential outcomes but it depends on them in a way that we completely control because within each strata, we have a completely randomized experiment.

So why would we want to stratify? Usually, it's just to improve power by reducing variance because the way you're going to analyze a stratified experiment is by looking first within strata where people, as far as you could tell, are quite as similar as you could make them.

They might share a gender. They might share some baseline characteristics. They might share the regions they are in, et cetera. Whatever you think is relevant to determine the potential outcome, you're going to make them similar. That means it's going to reduce the variance in the epsilon, whatever is left that you are not controlling, within each strata. And then you're going to aggregate the strata, and therefore, you're going to get an estimate that is more precise than if you had not stratified. Yeah?

AUDIENCE: How is stratification different from adding a dummy variable [INAUDIBLE]

ESTHER DUFLO: Ex-post? Of doing it ex-ante versus ex-post? That's a great question. So the quick answer is that if you do it ex-post, you have to estimate the effect of gender. So that's eating one degree of freedom.

If I do it ex-ante, whatever is the effect of gender doesn't matter, because I have controlled for it completely. So I don't have to estimate that effect. So if it's just one variable maybe it doesn't make a big difference. But when you create strata, you can create strata-- you can create very, very fine strata.

At the extreme, you could do pairwise randomization, which is randomized pairs, right, treatment and controlled pairs and just randomize one member of each pair as a treatment and control so people are really looking similar, as similar as you could make them. If you wanted to estimate-- if you wanted to just completely randomize and ex post control for all of the things that have made in the strata, that would be a lot of degrees of freedom to eat up. So that's the reason.

It's better to control. Then you don't have to worry about the estimation. One more thing to estimate is going to cost me some data. But if you haven't stratified, and for some reason, there is still there is imbalance at the end, because it could happen, you can always control.

So that's one thing that-- So that's why there is pretty much no downside to stratification. It can only help. So whenever possible, people tend to do it. Yeah?

AUDIENCE: Is there a situation in which you can [INAUDIBLE]?

ESTHER DUFLO: So that's a debated question. There have been papers about that. I am pretty sure that the jury is clearly came out and saying, no, you cannot overdo it. At worst, it's going to be irrelevant.

But it's never going to be-- it's never going to be bad, because you cannot do-- at worst, you're going to create strata that are irrelevant. You're back to a completely randomized design. You could do worse by adding tons of irrelevant strata instead of adding one relevant one, but in the same way that if you completely randomize, you would be worse than if you had one relevant one. So, yeah, so I think the bottom line is you can not overdo it.

With control variable ex-post, you can totally overdo it because you can eat up-- this is going to eat up all your degrees of freedom. So there is a tradeoff between adding more control variable is going to take away some of the sigma squared, but at the same time, you're paying it by degrees of freedom. With the stratification, there is no such tradeoff.

But it's actually not an obvious question. People have raised that maybe it would be an issue. It's not an issue.

So clustering is the opposite. Instead of randomizing, you've created buckets. Instead of randomizing within the buckets, you're randomizing across the buckets, which means that any bucket, if it's treated, everybody in the bucket is treated.

So, for example, the unit of randomization could be the school or the village or the family. So it will have power. It's always weakly worse than to do it at a lower level.

At the extreme, think of it where you were randomizing two regions. Effectively, your sample size is 2. When you do, it's a bit better than that because you have many people within the cluster, but not much better.

And, in fact, some people argue that the best way to analyze clustered experiment is just to treat each cluster as one observation. So you can see that then it could really hurt power a lot to group people into clusters and then randomize. So you never do it if you can avoid it.

But there are many cases where you cannot avoid it, either because the cluster is a natural unit of randomization-- for example, a school-level intervention, it happens in the school-- or because there are a lot of externalities. Anything you did to one person would immediately affect someone else. So the stable unit treatment value assumption would be biased if you didn't randomize at that group level.

Or sometimes it's just not practical because you cannot ask, for example, a nurse to treat people differently. Even if she could in principle, she just doesn't have the bandwidth to do it properly, and it wouldn't be done.

So these are the questions you have to ask yourself in how to randomize in the simple case where you can just randomize. And then there are cases where it seems impossible. Yeah?

AUDIENCE: [INAUDIBLE] simple randomization in terms of location in the sense that sometimes I know I need to stratify, but I don't know what my hidden variables are. So [INAUDIBLE] would that mean correcting--

ESTHER DUFLO: Oh, you mean to randomize, and then you don't like the result, you rerandomize?

AUDIENCE: No, just to look at what the variables could be that I could stratify for? I don't know the ages [INAUDIBLE] stratify--

ESTHER DUFLO: Yeah.

AUDIENCE: --age or gender, or-- is it your hypothesis or--

ESTHER DUFLO: Yeah, it's pure hypothesis. The beauty of it is that it doesn't matter, because if you get it wrong, it's no worse than fully randomized. So these are guesses which don't have a huge amount of-- it can help you.

AUDIENCE: True, but you can lose a lot of--

ESTHER DUFLO: You can lose power, yeah.

AUDIENCE: [INAUDIBLE] sometimes, stratifying is very expensive, right?

ESTHER DUFLO: Stratifying is not expensive per se, because, I mean, unless you need to go and collect the data to-- that you don't otherwise have to stratify. But otherwise, it just involved telling your software, when you randomize, make sure that you make sure that you first randomize among women and then among men. So per se, stratifying isn't costly.

The question I thought you were asking is, Can I-- suppose I have many variables and I don't really know how to construct my strata. I'm going to run one simple randomization, and then I look at the variable. It looks kind of ugly because it turns out that there are 70% of female in the treatment group, and I don't that. Can I do it again?

That's another question on which much ink has been spilled. And so my sense is that the jury is also out on this one, which is, actually, you can.

It's a little bit less good than stratifying if you knew what the strata were, because then it's very clear, and you can control for them. Here you're going to have an implicit stratification without having really-- without knowing exactly by what you've stratified. It depends what you've looked at. But you're still randomizing every time. And so unless you rerandomize, there is actually a theorem that Abhijit Banerjee and Sylvan Jason wrote, saying that unless you randomize more than the number of observations that you have, it's fine. You can just rerandomize.

So you can-- it introduces a tiny amount of bias. But compared to an enormous sample size, it doesn't really-- it's a small-- the intuition for it is that if you have many, many observations, and you randomize, say, four or five times, it still gives you very little choice. If you randomize so many times that you pick exactly the randomization you like, you might actually-- you're not randomizing anymore. You're choosing. But if you're doing a few times, you're choosing across such a small set that it doesn't matter too much.

OK, so suppose that you don't have a-- you're working with an NGO, and they say, it's not going to be possible to randomize for whatever reasons, and I'll go through the reasons. So first thing they could tell you is that we have money to cover 200 schools with our deworming program. We are going to work to cover 200 schools. And we cannot just go to see some schools and tell them that, sorry, we're making contact with you because we want to collect data, but we cannot randomize. They are not willing to do that.

So that happens, actually, less often than you might think, but it happens. And in that case, a design that's been very popular is phase in design, which is to say, OK, we are going to take these 200 schools, and we are going to randomly divide them into a few groups. And we're going to introduce the program progressively.

So, for example, you could say, well, in year 1, group A is going to get the treatment, and group B and C will be the comparison group for group A. In year 2, group A and B become treatment. Group C is the comparison. And in year 3, everybody is treated.

So, of course, in year 3, you don't have an experiment anymore. But you have an experiment in year 1 and in year 2. So that's been kind of a practical way of introducing randomization-- very popular at the beginning of doing randomization development. It has a number of caveats, actually, a number of problems.

The first one is that if the effect persists over time, it's a bit mushed up because suppose the effect takes 2 years to fully develop. Then, when you're comparing these guys, when you're looking at these guys, they are actually less treated than these guys. They are still treated but less than these guys.

And this one is still a comparison. And then, in the third year, you don't have everybody anymore. But people got treated in different ways. So you can deal with the analysis. But it's become a little bit messy.

Another thing that-- another problem with this kind of design is that typically, it's pretty clear what's going to happen. So you inform people of when they're going to be treated. And in some cases, the anticipation of being treated in the future could change your behavior.

So, for example, if you anticipate to receive a loan in the future because microcredit is coming to your village, you might decide that you're going to postpone some investment project until the loan comes, or, on the contrary, you're going to start a company, like a little business, because you know that you will have loans in the future.

So this has to be argued away in cases where you randomize with this design. But sometimes it's really the only way to go. Or you can just tell the NGO that's what you're going to do but not inform people that it's the exact sequence.

And so that's one way to proceed. It used to be very popular. For the two reason that I'm telling now is now people are trying to stay away from it when possible.

It doesn't mean that you don't treat everybody eventually if you think that the program is super effective. But you just don't make it a part of the design that is known and public and everybody is informed about. Yep?

AUDIENCE: I have a question about that. What if your NGO that wanted to do through their evaluation, they identify these three groups, but they say, let's say your intervention is a labor trafficking prevention program, and they identified group A as being more vulnerable. So they want to give the intervention to them first. Yeah.

ESTHER DUFLO: But then that's not randomized anymore. So that's not a-- maybe you can do something. Maybe you cannot. But it's not going to be in our city.

But let's see one-- let me pull this example to this one, which is actually a very useful one in this type of circumstances, where, say you're working with your NGOs. Easier for me to think about the credit bank, for example. But I think we could couch it in your example as well.

Suppose you're working with a bank, and you say, hey, we have this great idea. We're going to want to look at the marginal impact of giving money to people. So please randomize whether or not you approve credit or not.

They might look at you and say, you must be joking. There is a bunch of people I would never lend money to. So I'm not going to go for that.

So now you can say, well, yeah, sure, of course. I understand. But what are you using to score people? And then, typically, there is some kind of a score. So, for example, for banks, it's a credit score. Taking your example of trafficking, it might be about each community is identified by the fragility score or something like that. It might be more or less formal or informal. With bank and credit cards, it's usually quite formal.

And then, in the status quo, what they do is they say that anybody, say, in my example, above the score of 45 gets credit, and everybody below doesn't get credit. But it's not that these scores are perfect. In fact, many banks, if you talk to the manager, they will kind of fully acknowledge that these scores are a little bit black magic. And they're not very sure that they are so great. And maybe they could lend a little more, or maybe they should lend a little less.

So if there is such uncertainty, you can say, look, anybody below 30, no way. You just reject them all. Anybody above 60, they are great clients. You absolutely want to treat them. You're just going to treat them all.

But then, in that gray zone, between 30 and 60 in my particular example, maybe there is some scope for trying out. And, look, you're going to learn something that is useful for your business because if it turns out that the people right above the threshold actually are quite likely to default, then you might want to make the threshold higher in the future. On the contrary, if it turns out that these 30 to 45 guys are just as good, it turns out that you can lower the threshold in the future.

So in this example, they might be willing to randomize. They still prefer their good guys. So let's say the probability of treatment will be 85% and above threshold, and the probability of treatment will be 60% below threshold. But it could be even lower.

But the point is that you still have randomization there. And here you have two strata, one strata of above 45, one strata below 45, and that's your study sample.

So this randomization around the cutoff or randomization in the bubble has become quite popular because it works well with partners who can continue to do their business and get information precisely where there is a lot of uncertainty for them. And it is often the type of people we are interested in anyway because they are the type of people that would be exposed to an increase in this program if the program was made cheaper or it was expanded or something like that.

So that's also a useful design. A third design that comes very handy requires to understand the IV is encouragement design, which is particularly handy in situations where you have a program that's already there anyway.

For example, a nationwide program-- imagine a pension program is already there anyway. So it's not going to take it away from anybody. But even in those cases, the takeup for those programs might not be universal because there is a bunch of barriers that stand between people and actually getting the program-- for example, information or some loopholes-- sorry, the opposite of loopholes-- some hoops to jump and stuff like that.

So then you can say, well, I'm going to work with a bunch of activists-- take the pension program, I'm going to work with a bunch of activists-- I'm going to send them to identify eligible people for this program, for example, all people above the age of 65 who are poor and therefore are eligible for a pension program, and I'm going to help them apply if they are not currently getting it.

So then your sample is eligible people who are not currently getting the program. Let's say your activists are pretty effective. They get 45% of people to apply for the pension.

In the control group, stuff happen anywhere. Other people come, et cetera. 20% of people apply for the pension.

So the difference in 15% is randomly assigned, is due to your randomization. When you compare the treatment to the control group, that's not the effect of the pension. It's the effect of having an activist come and try to help you get the pension.

But by dividing by 15%-- 25%? By dividing by 25%, or multiplying by 4, you basically have an IV estimate of the effect of the pension. So the Ghana program that we studied last lecture is typically an encouragement design. The program is there anywhere. Everybody is entitled to go to secondary school. But by giving people a scholarship, I'm making it more likely.

Of course, in the analysis of any encouragement design, you're going to have to argue for the validity of your encouragement as an instrument, which means, A, you have the first stage. So here it's 45 minus 20. B, instrument is randomly assigned. That's usually true in an experiment. But C, it has no direct effect.

So in the case of the pension program, for example, you might think that a young activist coming to your house and helping you to get your pension is going to have to come three, four times anyway. Maybe there is a direct effect of this visit. You feel better. If you're an old person that nobody has ever visited in the last two, three months, and now this very fresh-faced young woman comes four or five times, you feel much better.

So in that case, you would worry about that. So maybe you have a home visit program as well in this group that you just don't make them apply to pension. So encouragement design requires one more level of thinking to make sure that your experiment will be valid. You have to do that thinking ex-ante because ex-post is too late to introduce the extra layer of treatment, et cetera. Yes?

AUDIENCE: Is it more common to come up with your own experiment that you want to test and then find a NGO that can do it, or to see a program that's already going, and you're like, oh, I [INAUDIBLE].

ESTHER DUFLO: There is a combination of everything. Sometimes, people come with something that they want to get evaluated. Sometimes, you have something you want to do anyway.

And then, usually, it's some combination of the two, which is people come to see you with this idea, and you say, fine. But they have in mind some very simple experiment. And then, when they've left your office after two hours, they have this ultracomplicated design in their hands that you've managed to sell to them because you want to do something interesting.

So it's kind of a combination. Basically, to run an experiment, you need a partner to implement the program. You need a research team to look at it. And you need money.

In principle, these three things need all to be there together. In principle, initiative could come from either three. It's actually pretty rare that money starts, partly because that creates partnerships that are not very effective, usually. But sometimes, the partner comes first. Sometimes, the research comes first.

OK, so this was sort of introducing randomization by stealth. And now I want to go over the second type of things I want to discuss about designs, which is, How do you design experiments to answer questions which are economic questions-- so not just does my program work, but an economic question?

And I'll give you two specific examples instead of talking in the abstract, one on estimating equilibrium effects of an intervention, and one is unpacking the effect of an intervention to understand better why it has the effect that it has.

The first question is-- the first example is a program that I worked on, research that I worked on in France. So all of Europe has high unemployment level for a long time. And governments are a little bit at a loss of what to do with this problem. So people are trying various things.

But one of the popular ways to help people is called the active labor market policy, which is when someone is unemployed, basically, doing a lot to try and help them get the job. So teach them to tie their shoelaces and put on a tie and show up on time at an appointment for a job interview, giving some phone calls to your contacts, et cetera. And so, in fact, many European governments give that job now to temp agencies that have sort of an employment placement bureau.

And there have been several randomization of such programs. And usually, the way they work is within a site-- for example, a town. Unemployed workers are assigned to one group or another. So they are assigned to the active labor market programs or they're not.

And those evaluations typically tend to find that people who are helped do better than people who are not helped. That's nice and well. But an important criticism against those evaluations is that the gains could be offset by displacement effect.

So suppose Leo gets helped, and I don't. Then he gets the one job that there is packing flowers. And the fact that he has it means I cannot get it.

So if we are going to compare Leo to me, we're going to find that Leo is doing better than me. So it's an experiment that is effective. In a sense, it's correct. But it is not a net effect, because it is just a musical chair.

There was only one job into the two of us. And the program gave him a little lead to get the job. So on the contrary, one could say, and one can write models where, actually, improving the productivity of the search effort actually improves the net amount of jobs that are available in the economy, for example, because firms just hate to search for workers.

And they prefer not even to post a vacancy if posting the vacancy is going to cost them the equivalent of half of a year of someone's salary to fill it up. So when there are too many unemployed people who have no idea what they are good at, that might actually reduce the number of vacancies that are actually filled. So it's an open question. And, of course, it could be a little bit of both.

So these are what's called equilibrium effect, which means, in equilibrium, how many jobs are there is going to determine whether or not this is extra unemployment or not. So when you do that, well, the only way to do this, really, is to randomize into steps in order to find out whether-- take the example of Leo versus me-- the fact that there are many Leos in the labor market that are helped and I'm not, does it make me worse off compared to my situation if I were into an entire other town on this side of the room where the program doesn't exist?

So the randomization design here is to go in two steps. Two steps are done at one time. But you first randomly assign the proportion of treated to areas, and then you randomly assign treatment status to individual within areas. Let me show you a graph of that.

So basically, what we did is we created-- these type of things can only be done at a fairly large scale because you need the skill to get the equilibrium. So in France, people don't like to move very much. So a town is roughly a labor market. People aren't going to move to get another job.

So what we did is we took-- we worked with an employment agency in about half of France. And we divided the towns into a strata of five. So literally, it's fully stratified in that sense.

So in each strata, which are quintuplets, we have five towns. And in one town now in each of these strata, we picked randomly one town where we treated nobody, one town where we treated-- sorry, in all of ideas-- backtrack. In all of the quintuplet, we took one town where we treat nobody, one town where we treated 25% of people, one town where we treated 50% of people, one town where we treated 75% people, and one town where we treated everybody.

In retrospect, that everybody one wasn't that useful because it's not everybody. It's everybody who is eligible for this program, which are the young unemployed workers. So to start with, they're just they're a fraction of the people. But anyway, that's how we designed it.

And now you can look at the effect of being helped within one region. So the Leo versus me comparison, which is what people have done traditionally, by comparing Mr. Blue here to Mr. White here, Mr. Blue, of course, within each town, once we determine the proportion, the people to be selected are randomly assigned, OK?

So we can do this Leo versus me. And then we can also do me versus Frankie over there, who is actually in a town where nobody got treated. And that's going to answer the question of whether the fact that Leo was treated hurts me by looking at what is the situation of someone-- what is the chance of someone to get a job back within six months if nobody around them got this program versus some other people got this program.

So that's the idea for the program. I think everything is here. Yeah, so the program is-- the target population is young unemployed worker, so people age less than 30 years old and unemployed for more than six months but with some college.

So what we have is then these people in the all-white area, we call them super control group. So they are individually assigned 0% area. By comparing the assigned to control and assigned to super control, we get the displacement effect by comparing assigned to treatment. And the super control, we get the full effect of the treated of the program. By comparing assigned to treatment to control, we get the effect within. So that's the potential job-stealing effect.

The equation is here. This regression is run within places where-- so this is controlling for the dc. So it's controlling for the area level. So it is basically a comparison of Leo versus me inside one labor market.

And what it shows is that very little effect for women-- this is the effect-- the question that is being asked is not very well labeled. The question has been asked is, Have you found a job-- the outcome is here-- have you found a job of-- a long-term duration job, so more than six months-- have you found a job for more than six months within six months?

So six months after the program, I'm looking at you again. And I ask you whether you have in your hand a contract for an employment for more than six months. And we find that about 17% of control people have found a job anyway. But there is a pretty large effect of men of 5 percentage point.

So this is the within effect. And now I'm going to add the comparison of the control to the super control. So this 5.1% here is still there. It's the comparison of Leo and me. But now I'm comparing me and Frankie.

And you see that I'm hurt, actually. There is a negative effect of being in this area compared to not being in this area of 3.9 percentage point. And then the net effect would be then to compare him to her.

And the net effect on him is really not that large. It's 1.2%. But it looks big in the treatment area because he's taken my job. So I'm doing worse. He's not doing that much better. But if we did it within, we would think that there is a large effect.

So this idea of randomizing into steps is going to be useful whenever you think that there are spillovers and you're interested in the spillovers. Then you can think of it in your mind as one treatment is being directly treated and one treatment is being exposed to people who are treated. So going back to the design, you have a combination of cluster-level randomization because you randomize a cluster into a fraction treated, and an individual-level randomization because within each cluster, you're going to randomize who are actually treated.

AUDIENCE: Can I just revisit the results to see if I understand? So 5% for men, that's considering displacement-- or not considering displacement, [INAUDIBLE] displacement effects are happening.

ESTHER DUFLO: Exactly. It's the difference between, within a labor market, the treated and untreated. So it's a difference of you versus me.

AUDIENCE: So if we were to generalize, then, the success of these programs in the literature is spurious, given that--

ESTHER DUFLO: That would be the conclusion of that, yeah. We find that what looks successful is just the fact that within an area, the people who are helped are somewhat more likely to get a job at the expense of people who are not helped.

AUDIENCE: That sounds like a very important result. How would a government interpret that, or would you do the study again in a different country? Or are there any other variables that are [INAUDIBLE]?

ESTHER DUFLO: I can tell you how they interpreted this result, which is not one of my success in term of policy influence. So soon after we came up with this result, the program was scaled up. But that's not how they should do it. You should think of that as being pretty bad.

So there actually-- there is a twist to it, though, just in terms of the substance, is that it turns out the evaluation spanned the recession that followed the 2008 crisis. And you can look separately at places that were most affected by the recession, people who are less affected by the recession, before and after.

And what you find is that the displacement effect is the strongest in recession time. So in recession time, presumably, where firms are not hiring anywhere-- so you can help the search effort as much as you want. It's not going to be all that helpful.

In a nonrecession time, in good times, then the displacement effect is much smaller. So in that case, in nonrecession times, firms would be hiring if they found capable people around. And the fact that you are training some people into presenting themselves better and retraining for some skills, et cetera, helps.

Another thing that you're learning from this experiment, and it's kind of always-- is that-- so this is the evaluation of this particular program. But if you're willing to be a bit-- willing to use models to go one step further, you can say, generally, it suggests that means that search effort, the productivity of search, is largely competition between workers.

That means in particular during recession times, which means that maybe we need not be too worried about unemployment benefits. The problem of unemployment insurance is that it makes people not that enthusiastic of looking for a job. But that's only a problem if they would find a job should they search for one.

So in recession time, because of social insurance concerns, you might want to increase the length of unemployment insurance because that's when people really need it. And what this type of result suggests is that in addition, they probably will not be too much of an efficiency cost of increasing the length of unemployment insurance in recession time because, yes, everybody is going to search a little less. But it's not that there are jobs for them to find anyway. So the jobs that there are are going to be filled regardless with the existing search effort.

So one lesson that you can take that's a little broader than just this is that in recession time, both the benefits of increasing UI is high because of insurance motives, and maybe the cost is not that high. So that would give you a sort of an argument for doing this, modulating the length of unemployment insurance with the macroeconomic cycle.

So that's sort of-- and beyond this particular topic, this idea of randomizing in two steps has been used recently to look at various effects of spillover, any kind of spillover. Here they are coming from equilibrium effects, but it could be from adoption, contagion, et cetera, can be looked at with similar designs.

The last example I want to give you today is-- also a very popular way to do things is this kind of answering a little bit Joseph's question is that-- who comes up with the program? Is it the implementing partner, and then you're kind of evaluating whatever they are interested in? Or is it the researcher?

And often, actually, it's some combination of the two. And I'll give you one very nice example from a paper by Abhijit Banerjee and Ben Olken, who are both here, and Rema Hanna, who is at the Kennedy School.

And they're looking at the program in Indonesia called the Raskin program. The Raskin program provides eligible households, which are poor households, with 15 kilogram per month of heavily subsidized rice. And that's a pretty terrible program. It's full of corruption. Many, many rice gets disappeared, et cetera. It goes to the wrong people.

There are many reasons for that. But one of the reasons is that-- but for whatever reason, the government likes this program. They're not going to get rid of it. So that's not on the table. But they want to try and make it better.

And right now, one problem that the program has is that the information among citizens is low. So a survey that they did suggests that only 30% of eligible households know that they are Raskin-eligible. And the beneficiaries also believ that the copay is 25% higher than it actually is, probably because someone helpfully informed them of that somewhat higher number. And so as a result, the eligible people for the subsidy only received about a third of the intended subsidy between the fact that some of them don't get anything and some of them overpay for the rice.

So the hypothesis of the government is that this level of misinformation may give officials an advantage in bargaining with the villagers and may lead them to basically take too much from them. And they want to try and improve program transparency.

So they set up this randomized trial in 572 villages. And in 378 of them-- so randomized at the village level-- people would get a card which informed them that they are eligible and for how much rice. So that's kind of the card they are getting. So they have their names, saying that you are eligible, and then how much rice you're eligible to get.

So that's the basic thing of the card-- very simple. The government is interested in distributing this thing.

So, basically, as far as I understand the genesis of this project, the government comes to them and say, hey, we are interested in distributing this card. But we want to know whether it's going to improve the program. Can you help us?

And then they say, sure. But let's do a little more. Suppose that the card do work. What else might you want to know?

You might want to know whether it's the information that's in the card, whether more information would make it work better-- so whether people pay attention to what's on the card, whether the physical card is important or a list published in the village is sufficient, whether it's going to be accountability effect-- so now that there is a card, the village officials feel that they could go to jail if they cheat because they think that there is accountability. So can you improve this sense, increase this sense of accountability? So all of these questions, you might want to ask.

So this leads you to-- if you want to answer all of them, this leads you to a sort of pizza pie of design, which is what I'm going to show now. So the first thing is public versus private information. So what matters is-- because when I give you a card, I give you a card. So you're informed. But the village official also know that everybody has a card. And I, as a card holder, know that the village official know that I have a card.

And then she knows that I know that I know that she knows. So all of these things will also enter in our decision process. So the question is whether this public information, this second level of information-- not only my own information, but the fact that everybody else is informed-- also matters.

So what I did here is that in some villages, they just give people their own card, and the village had got one list, and one copy was posted. And in the public information treatment-- so this is randomized at the village level as well-- in the public information treatment, they printed many copies of the list all over the village. And they printed these posters. So now everybody is informed. So that's kind of one extra treatment.

Another thing you might wonder is whether the precise information that's on the card matter or it's just that it's a card. So it now looks a little bit more official. So here, do you see the difference? I guess you may have-- don't read the English here, but look at the two cards. Do you see the difference between the two cards?

AUDIENCE: Second bullet point.

ESTHER DUFLO: Yeah, see the second bullet point here. And even knowing no Indonesian, no Bahasa Indonesian, you could guess that it tells them that the price is 1,600 rupiahs per kilo of rice. So one doesn't have the price. One has the price. So now we can look at whether people who got the card with the price end up paying a lower price than people who didn't.

The second thing is you can see whether the physical cards matter or just the information matters. So in all villages, the full list of eligible beneficiaries was distributed. But they varied whether they send the card to everybody or they send the card to only the bottom 10% of the population-- so everybody who is informed, everybody is in the list. So they can know that they are in the list. But only the poorest get the card. So this is designed to test the physical role of the card in bargaining.

And finally, the accountability one-- so it's to test the idea that the card implies checking in on the official. So they had two versions of the card. One was this one, and one was this one with this coupon. So every time you get rice, you're supposed to give the guy a coupon.

And then the coupon was supposed to give it to the top. In reality, nothing was ever done with those coupons. But there was implicit in all of that the possibility that someone is checking, there is this bar code scan, et cetera.

So what it gives you is that you want to run four different dimensions on 4-- so 16 possibilities, public versus private information, information on the card, who received the card, and the coupons. So it gives you this big matrix, a factorial matrix. Or in principle, it's completely randomized such that everybody is in.

Now you have to ask yourself, when you have a design that's complicated like that, first, you have to ask yourself whether-- am I going to be able to keep my decks aligned, and am I going to get terribly confused when this is actually done? So if Ben Olken is involved in something, you can make sure that-- you can be sure that his ducks will be aligned. So he had no problem doing that straight.

But another issue is whether or not-- how to think about your sample size because you could think of your sample size one of two ways. You could think, well, the only thing I'm interested in-- I'm doing this big pizza pie of design, but I'm not interested in separating each slice from each other slice, each box from each other box. I'm just interested in having enough power to compare price versus no price.

So I'm going to aggregate these eight. And comparing them to those eight, I'm going to-- or is it that I'm actually interested in the interactions between the two? And your power will depend-- your sample size will depend on that.

So a big mistake people make in designing experiments is the power of their experiment for treatment versus control. And then they have lots and lots of interesting subtreatments, but there is no power for any of that, because they didn't think of powering the experiment for the subtreatments. And therefore, it's not as informative as it could be.

So on the other hand, if you have a complicated design with a lot of subtreatments, the sample design could become "giganormous" before you're done if you were interested in separating every single box to every single box. So in this case, I can tell you what they are powered for. They are powered for any two-by-two comparisons-- so price versus no price, coupons versus no coupons, card versus all versus card to bottom 10, but not any of the interaction.

So basically, they commit ex-ante not really to look at the interactions. Then you can ex-post decide to pick. But that's-- the data comes from followup survey conducted 2 months, 8 months, 18 months after the cards. And they also interviewed the village leaders and the other sample beneficiaries because, of course, they are the ones where most of the action is going on.

OK, so here is the main effect. So first of all, who gets the card? So this is what you get. In real experiments, things don't go exactly-- this is a real program. So ideally, we want this to be 100%. But only 28% of people actually received the card.

14% ever used the card. 14% ever used the card. 6% of the control group claimed to have received the card anyway. 9% correctly-- they could have moved or it could be a mistake-- 9% correctly identified-- 9% more correctly identified their own status, so from 30% to 39%.

By the way, in most reporting of randomized controlled trials, you report the control mean in the bottom. It's conventional because that way you can see-- you can compare the effect size to the mean. The ineligible households, some people got their hands on a card anyway, but not too many.

So that often happens in the real world where it doesn't work as well as you can. So the results suggest that the cards had an impact, though. Oops, I wanted to show you the-- I haven't shown you the-- I should have shown you a main effect, a table of main effect here.

When I cut down some of the slides, I cut down the most important slide, which is what is the effect of receiving the cards on the subsidy. So I'll tell you the results, and then we'll look at a lot of subresults.

The result is that it makes it more likely that you get rice. You pay less for it. So on average, people get a higher level of subsidy, about a third more, a quarter more. So the increase in subsidy in total received by the eligible was 25%.

This was cost-effective because it's really cheap to provide the card. So in term here, in term of policy, the government, finding this out, expanded the card. And now what we can look at is to investigate the mechanism by comparing all of the-- by comparing all of the set sample.

So first, let's look at public information. So in the public information case, people were more likely to have seen the list. They were 14% more likely to have seen the list in public info and only 2% more likely to have seen the list in standard info.

They also believed that other people have seen the list. So they think that in the public information, they think that a lot of people have seen the list because, of course, they have seen the list everywhere. So there is an effect on your perception of other people knowing what's going on.

And when we look at the impact, we are now comparing the public info-- one line for public info, one line for standard info, which is just the card. And you can see that in the public info compared to the control, they get 9,000 more in subsidies. And in the standard info, it was about 5,000 more in subsidies, which is a combination of purchasing more rice and paying less for it.

So the public information seems to increase how much subsidy people get. The price information-- remember the cards, and we can look at it. So in the price information, people do get more rice, and they get it at a lower price. So even though it's just this one line, this one line actually makes a difference, suggesting that the actual information that's on the card also matters.

And then, finally, does it matter to receive the card yourself? So this is the-- we should focus on the other eligible household, the one who are not receiving the cards when not everybody gets it. You see that when they get a card, their subsidy increased by 5,000 rupiahs, more or less, but when they don't get the card, by only 1,600. So the physical card also matters.

I didn't print the table for accountability. But the accountability didn't do much. So basically, in all of these treatments, we are learning that from these little cards, actually, the little card, it doesn't look like it, but it's really a bundle of stuff-- information that you receive, the specific type of information, the ability to-- the physical card, the fact that you know that other people got the card.

And this experiment managed to kind of create all of this treatment to unbundle these effects. And what they find is that in the end, all of these matters, except the accountability effect that doesn't seem to be driving much of the results.