In this presentation we’re going to talk about randomized experiments. We’ve already talked about several of the issues related to this design. Randomized experiments are often considered the quote, unquote gold standard of research designs as they allow us to talk about impacts.
A bit of review; it’s worth saying things often. When we speak about an impact we’re talking about a change in our outcomes that would not have occurred had the program been absent, so we often use this language of A causes B, but what we really mean is if A, our program is introduced, B, the outcome is more likely to result than if we do not introduce a program. So that’s the setup of what we’re after and we’re going to talk further about how randomized experiments, randomized control trials, these types of designs can help us explore impact.
So when is it appropriate to conduct an impact evaluation? There are several points when we might consider looking at impacts. We can think about doing it in a pilot stage, so after the development of a program and we’re putting it out in the field and we’re wanting to see what impacts the intervention has, we can consider a randomized experiment there, and they’re also useful in an ongoing basis potentially so when a major change occurs in a program we can think about doing an impact evaluation. We could look at a program that’s been in existence for a long time and maybe hasn’t been examined for impacts in a while and look to see if the intervention is as powerful as it was shown in the past. So there’s several points in the life course of an intervention where we might consider an impact evaluation, and we have to think about in our minds what are some prerequisites for us to be able to make meaningful assessments of program impacts. It really revolves around earlier forms of evaluation, so having done good work and specifying the expected effects, back again to the theory of treatment, and we need to be able to have a strong assumption of the effects that can be produced by the program and be plausible and we must have measurable outcomes. So these are fairly basic prerequisites but it’s incumbent upon us to meet those prerequisites.
So remember in its simplest form when we speak of a randomized experiment what we’re talking about is let’s say we have a group of people that are eligible for treatment, that are eligible to participate in our intervention and we use a random mechanism to assign folks to either receive the treatment or to go into a control condition that may be some other treatment; it may be business as usual. But the key to finding elements of this design is that the receipt of treatment is purely and only driven by this random mechanism that assigns folks to one group or the other. So the question is why does that work, and hopefully through a lot of these presentations we’ve kind of hit on why this might work. And the reason it works is it removes all rival alternative explanations for the observed effects. All those alternative causes that could potentially look like the treatment are removed by design, so all those... not all, most of those threats to validity are removed and the reason it does this is that the selection mechanism is known and we can model it. So the only way that folks receive treatment, at least in the general sense, is that they were randomly assigned to it; there is nothing about them that is related to the receipt of that treatment, so their gender, their race or ethnicity, their prior academic achievement, none of those things are related to the mechanism of how a person receives treatment or the control condition, and so in essence, assuming we had a large enough group of people, the expected values of all variables, whether we’ve observed them or they’re unobserved to us, are expected to be equivalent in the two groups, and that’s where the power of this design comes is that although most of the confounding threats to validity that we’ve spoken about are removed by this design because they’re not there anymore. The only reason somebody got into a group was by random assignment.

Logic of Randomized Experiments

- Random assignment reduces the plausibility of alternative explanations for observed effects
- Why RA works
  - Removes confounds of alt. causes with treatment condition
  - Reduces plausibility of threats to validity
  - Equating of groups on expected values of all variables
  - Selection mechanism is known and can be modeled
To talk about random assignment and internal threats to validity a little more and go back to kind of those three conditions we talked about earlier that lead to causal relationships, the design itself ensures temporal ordering so that they cause precedes the effects, and do that by design; you randomly assign folks to condition, you then implement the treatment and then you observe the outcomes. And as I said in the previous slide we know the mechanism so that prevents selection by definition; it’s a random process; there’s no attributes about folks that lead to a higher likelihood to receive treatment or not that can also be related to their outcomes at the end of treatment, and that’s selection bias. And then it renders most alternative explanations implausible, those threats to validity by design through this equating process of the random assignment. It reduces the likelihood of most of the threats confounding with the treatment. We still have to be wary of a few and we’ll discuss them in later slides, but in essence you’ve eliminated these threats to validity by design.
Often randomized control trials or randomized experiments bring up questions about the ethics of randomly assigning people to receive treatments. It can be quite common when one presents the possibility or using random assignment to evaluate the outcomes of a program, and those are understandable concerns, but often randomized experiments are quite ethical to do. There are some general rules of thumb or heuristics you might use to think about; is it appropriate and ethical to use randomized experiments. And first present practice must be in need of improvement, setting the groundwork in the case where something needs to happen, something needs to change from current practice. Often we don’t know whether interventions are actually effective, whether they actually have positive impacts on kids in education, so that’s a compelling and ethical reason to examine them with arguably the strongest design, to be able to determine does this intervention actually have an impact, and we need to think too, the stronger that case can be made there’s potentially a stronger case to be made for actually having folks spend resources to implement it which also has some ethical questions around it; how do we make decisions and how to use limited resources. Often we... there should be no simpler alternatives for evaluating the intervention; maybe there’s other ways. We can do an evaluation to fulfil the goals that are required by stakeholders. The results should be potentially important for policy so we should have a justification for making the case that we need to know the impact of this intervention on outcomes, and then ultimately we need to design it in a way that is ethical, that conforms to all the requirements of human subjects research, and also being aware and attuned to the
needs of service providers or organizations where we might want to implement this
design. Those are often the same things but sometimes the decision points are
slightly different so we need to consider those two angles and how we think about
the ethics of randomized experiments. Ultimately I think the case can be made quite
strongly of why these are ethical. Often it comes down to limited resources. In many
senses random assignment is probably the most ethical way of apportioning scarce
resources because it’s not related to any characteristics of folks and so we can avoid
actually ethical dilemmas by using a random mechanism.
Research Questions and Theory

- RE are, by design, better suited for some types of questions
  - What works?
  - Who does it work for?
  - How long does it work?

  - “Does participation in program x lead to higher outcomes y than would have been the case in business as usual?”

Slide 7
So by design randomized experiments are better suited for some sorts of questions than others, and the types of questions that this design works best for is precisely that; what works. What is the most effective treatment? We can also get at questions of who does it work for, is it differentially effective for different types of people and participants and how long does it work for? So all of these elements are manipulable of which a randomized experiment is intimately able to do. And the types of question if we were to write this up as a question, does participation in program X lead to higher outcomes Y than would have been the case in “business as usual?” That’s a nonspecific but general type of question that randomized experiment is best suited to answer; it’s very clear, it gives us the conditions of the randomized experiment of program X versus business as usual, it makes hypotheses about the direction, our expected direction of results. And this kind of question screams, if you will, randomized experiment.
So just as by design a randomized experiment is suited to answer the what works types of questions because of its strong internal validity, they’re not well suited for other types of questions; they’re not useful to us for elucidating the contours of the problem or giving definition to a problem, they’re not useful for fidelity of treatment and programmatic planning. But we also need to remember that those types of questions – fidelity, definition of problem, those are important elements of a well-designed randomized experiment and they help us to understand the results of a randomized experiment.
Although randomized experiments may seem rather simple it’s important that we consider the vital role that theory plays in the design of a good randomized experiment. Theory underlies how we specify the conditions of what is treatment; that should all have a theoretical framework, positioning it and writing our hypothesis about why it might work and what it looks like. Theory also tells us what the comparison condition should be and we can use that to then design those conditions in the experiment. Theory helps us identify what is the actual unit at which the implementation will take place; is it a school-level random assignment where the intervention is provided to everyone within a school and we randomize schools to treatment or the comparison business as usual control conditions or not, or do we at a classroom level or do it at a head level, student level. Theory should undergird our decision specifying that part of the design. It will help us think about how we go about analyzing the data that comes out of the experiment, it helps us to think about and then create systems to capture contextual variables – what are those that we really need to pay attention to, need to think about. That could be some of these potential, historical threats to validity, those types of things, and other important contextual variables that may moderate or mediate the effects, and ultimately theory helps us interpret what we find and to what extent do the findings conform to the hypotheses that we made from theory or not. So theory is crucial to specifying and designing a good experiment and we should always ground our designs in that theory.
We’re now going to talk about some specific elements that should be considered when thinking about randomized experiments. They’re not unique necessarily to randomized experiments; we should be thinking about these issues in all of our designs, but they do play a role in understanding and designing randomized experiments. The first one is what is the target population? That can come from a theoretical basis of what the target population is for the intervention based on the stated problem and the theory of treatment underlying the intervention, and that means we have to pay attention to the eligibility criteria. Who is eligible to participate in the study, who is eligible to get to the point to then be randomly assigned to treatment or control? There are valid reasons for determining some folks eligible and some folks ineligible. We need to think around who is excluded, who is explicitly excluded by eligibility criteria, who is excluded as a function of the way we’ve set up the target population, so maybe it’s a certain grade level and maybe it’s a certain ethnic group depending on our study, various things like that, so whether we’re explicit about it or just by the nature one group gets excluded. We need to consider that and we need to think why is that? Well specifically in the context of randomized experiments they have strong internal validity but the external validity, the ability to generalize from this one study in this one specific setting of this one group of students or participants is more limited. So thinking about those eligibility criteria helps us to better understand who it works for perhaps. In thinking about can we generalize this study, these findings from a study with a strong internal validity, can we generalize it to another group of students. So thinking through those elements are
critical to understand our studies.
We need to think also about the level of analysis or the level at which treatment is applied, where are we implementing the treatment and the unit of analysis is not always the individual. If we’ve randomly assigned treatment to schools let’s say, the level of analysis is the school. And a lot of that is driven by program theory or very practical considerations of how to do the work. This looks back to some of the other threats to validity. So let’s think about if we had a teacher-level intervention—we’re going to train teachers to do some type of reading program—in a simple context we might think of randomly assigning teachers to conditions within schools. Well that may work but it also may open the door to some threats to validity of crossover or contamination between the treatment and control group because what do we know about teachers? They talk to each other, and so if we have teachers in two separate conditions in the same school teaching the same subject we’ve opened the door to some of our treatment characteristics running over into our control condition, so we have to be careful of that. So maybe we instead of randomly assigning teachers we randomly assign schools to implement the treatment. So that’s fine, and then we have one school that’s all treatment; all the teachers are implementing, and another that is all control. And so the logic of the experiment is valid at the school level or at the teacher level in this case. Either one is valid, but what we must consider is that the cost and the difficulty level is increased as we go higher, so in order to appropriately power the study we’re going to have to enroll more schools which has a higher cost.
We also need to pay special attention to small sample sizes. Random assignments work best, if you will, in larger samples of participants and that’s because the more people that we have the stronger is our belief is that those observed and unobserved differences between people will be balanced out by the random process between groups. In a small sample of participants just random variation can unbalance the assignment, so you could think of a world where say we had ten students of two ethnic groups. We could easily randomly assign and get all of one ethnic group into one group and all of the other ethnic group into another. That’s much less likely to happen; the more participants we have... when we have small samples that can most certainly happen – it’s called “unhappy assignment,” and we want to avoid that if at all possible so we’re better off having a sufficient number of people and trying to avoid small samples; randomized experiment is not the best, I would argue, in those cases. Smaller samples, we can think about maxing or blocking so these are terms... In the example I just gave we could randomly assign within ethnic groups, we could pull out ethnic group one, randomly assign a treatment and control; ethnic group two randomly assign to treatment and control and that can help us avoid the unhappy assignment. And then we could just do it again, re-randomize; still fine. We could create all possible randomizations, discard the unpleasant ones, then randomly choose from the remaining. So there are strategies I would argue that in the case of small samples that we might want to redo the random assignment study.
Treatment implementation and fidelity of implementation are crucial to understanding impact evaluations using random assignment and all impact evaluations. So we have to know what is the treatment, what does it mean to participate, and not just that; we also need to understand the control condition, so we say business as usual but what does that actually mean? And the way to think about it is the control condition is not the absence of treatment; it is a treatment in and of itself, and just as we spend time measuring the fidelity of implementation of our intervention, the treatment, we should also consider what does that mean in the business as usual control condition? We should measure fidelity as well and come up with a way of gauging what is actually happening there that gives us the ability to talk about the treatment contrast and to understand the magnitude of the impact. Or more importantly, if we don’t find an impact it helps us to better understand why we might not have found an impact.

<table>
<thead>
<tr>
<th>Interventions and Methods for their Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Tx specification</td>
</tr>
<tr>
<td>– What is the treatment?</td>
</tr>
<tr>
<td>– What does ‘program participation’ mean?</td>
</tr>
<tr>
<td>• Fidelity of treatment implementation</td>
</tr>
<tr>
<td>• Control condition specification</td>
</tr>
<tr>
<td>– “Business as usual”</td>
</tr>
<tr>
<td>– The control is a treatment</td>
</tr>
<tr>
<td>– Attention to control just as important as the treatment of interest</td>
</tr>
</tbody>
</table>
So the critical elements of the randomized experiment in this design is that we have a control group that is equivalent to the treatment group and as one of the readings talks about the equivalence is in terms of the composition so we have the same mix of persons in both groups; that’s how we get... that’s the leverage whereby we’re getting around these threats to validity. They should also have identical predispositions; that should be balanced across the group so folks that are likely to respond to intervention are equally likely in both groups and also conversely folks that are not likely to respond are equally likely to be present in both, and identical experiences. So these are those time-related processes, maturation and history elements, those threats to validity, they should be identical with the exception of the receipt of treatment. And so when we meet those equivalence conditions then we can be reasonably assured that the random assignment is allowing us to get to discussion of impacts.
So it’s important that the actual method of random assignment is systematic and we want the probability of being randomly assigned to be equivalent across units of the target population so an individual has an equal probability of going into either group. We’ve covered this; we need a certain number of units to do it properly and perhaps more importantly we need to think about how is the assignment done, is it done in a way that cannot be anticipated or subverted or manipulated in any way by potential treatment individuals, or non-treatment individuals. We don’t want people to be able to get around the random assignments and advocate for themselves to get into a treatment versus control. There’s ways that we can do that by thinking about when we time the random assignment and when folks know that they have been randomly assigned. There’s a good example of that with the Tennessee Star experiment which is a class size randomized control trial that was done in the late ‘80s in Tennessee and randomized as close as possible to the beginning of the school year under a worry that people might be able to subvert the random assignment; they randomized kids to smaller class size or regular class size, and they were afraid that parents might advocate to get their children into the smaller class size and that might be related to social class which could be related to outcomes. So you want to avoid that if at all possible. You need to think about who actually controls the assignment procedure, who’s actually doing it. As I mentioned before when does it take place and how is it structured, so we don’t actually have to assign whole numbers of folks to the two groups, we can think about unbalanced allocation as well. So perhaps the interventions are very costly so we can find could assign fewer to the treatment and

Method of Random Assignment

- Method of randomization is systematic
  - All units in the target population must have the same probability of selection into either group
  - Requires an appropriate number of units to approach statistical equivalence
  - Assignment cannot be anticipated, subverted, or manipulated

- Who controls the assignment procedure?

- When does the assignment take place?
  - Should take place close to beginning of Tx

- How is the assignment structured?
  - Blocking
  - Balanced versus unbalanced
more to the control or vice versa; we could do the same.
So when we think about analyzing data from randomized experiments the basic core analysis is quite straightforward almost to the point of being that simple treatment/control contrast, under the assumption of proper randomization. We also have to be mindful of attrition. I believe we talked about this in the threats to validity so we don’t have differential attrition of folks between two different groups or other non-response threats; that can make the analysis less than straightforward and as a phrase, “analyze them as you have randomized them.” We need to also think about quality assurance, so was the treatment done with fidelity; if it has not can we reasonably say that there’s actually been a treatment, so that’s why we attend to fidelity of implementation. We need to come up with plans for making sure that folks stay in their assignments and that a person in the control group does not somehow surreptitiously receive treatment; it can happen. So we need to tend to all those things as we think about analysis and that means we have to think about them before we implement them and have plans for collecting that data or maintaining the integrity of our treatment assignments. All of that will then lead to our ability to go beyond the treatment contrast, to go deeper perhaps. It helps us to understand no-effect findings which do happen and we need to understand why there’s no effect, and potentially differential effects.
In summary we've done the barest amount here which are readings, the Shadich, Cook and Campbell book as well as the Rossi and Lipsey and Freeman book go into more detail. A randomized control trial, a randomized experiment, are the gold standard for looking at impacts. They have the strongest internal validity which is what we want to be able to say; this treatment caused this effect, this reading program led to this size of reading gain among students, but what we really have to realize is that that's in theory the gold standard. If we have we have a high quality design done with due diligence and planned all the quality assurance in, so attending to fidelity, attending to maintaining the integrity of the treatment assignment, heading off attrition or differential attrition. If we’ve done these things then yes it is the gold standard, but often randomized experiments do not weigh out as cleanly as they are designed and so we need to be mindful of that and worried about especially differential attrition among treatment and control, and we also have to realize they’re not appropriate for all questions and contexts; there are times when it’s not appropriate to do a randomized experiment given your questions and given the context in which you’re working. Sometimes that’s a more mechanical reason; it’s just not possible. Others, there may be ethical concerns about randomized assignment that prevent one from using it, and we also need to remember that they can be very costly; they can take a long time to do and do well, and those reasons alone may disallow us from using the design.