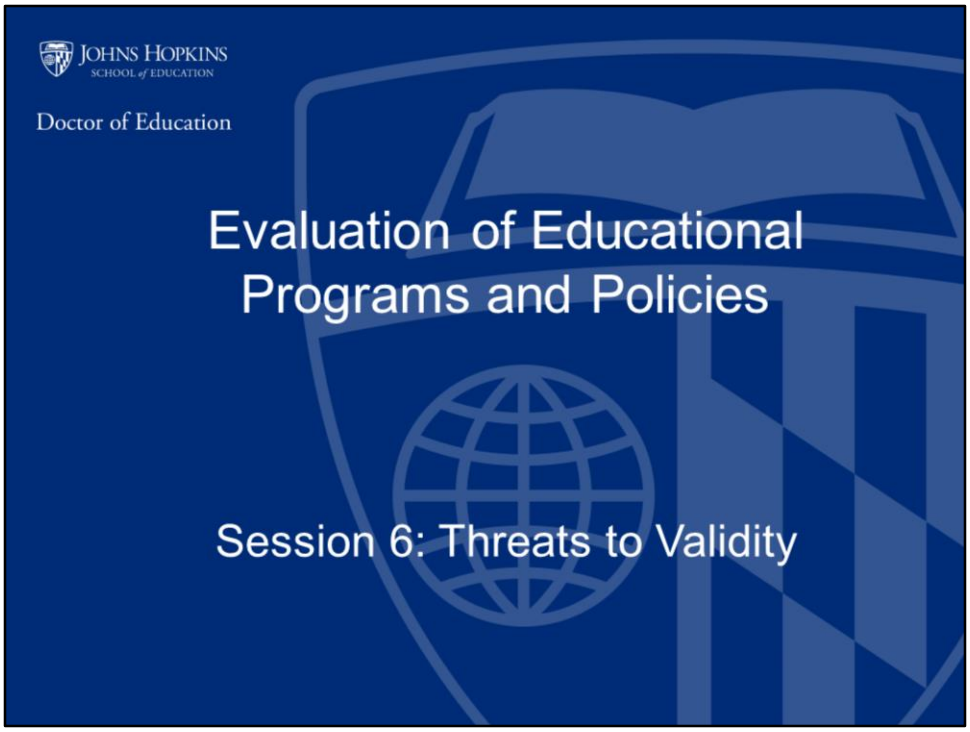
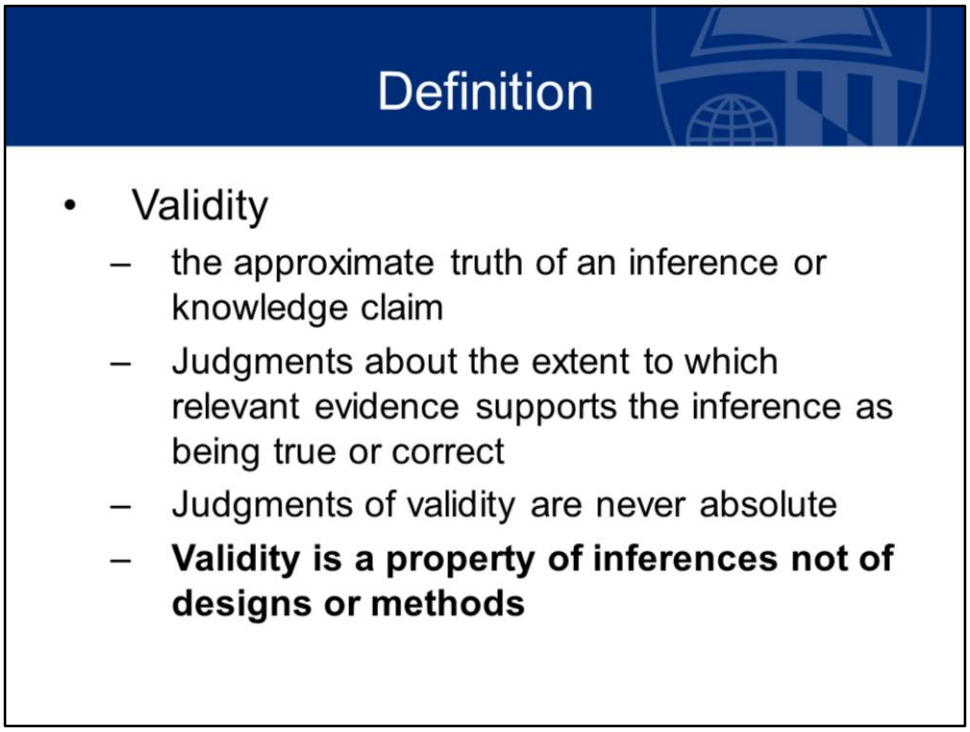
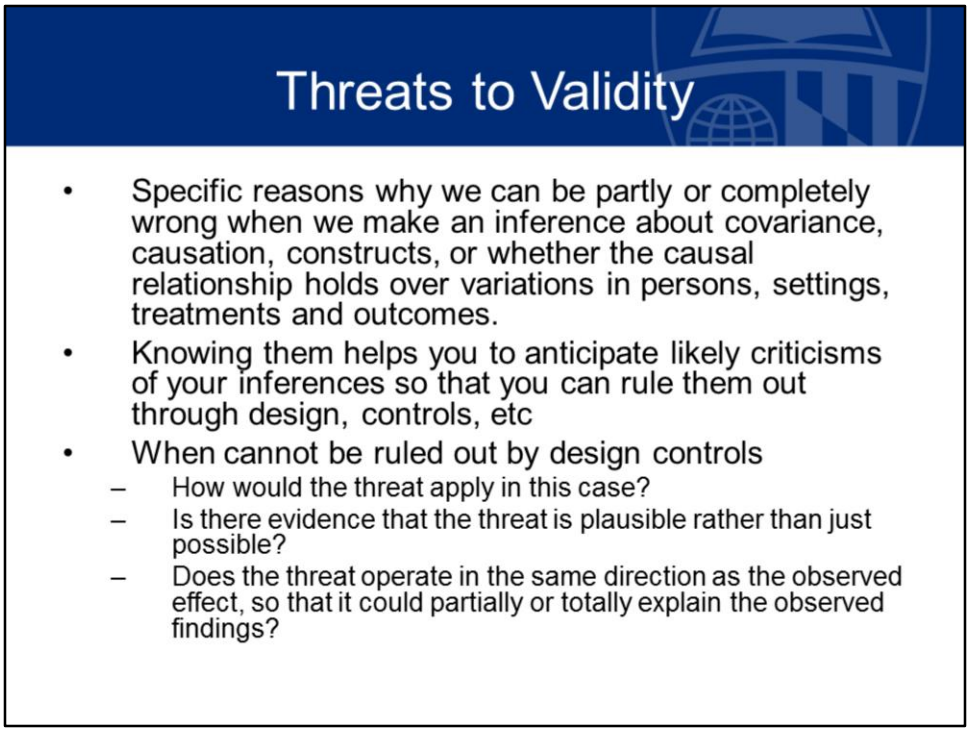


In this presentation we're going to talk about threats to validity and by validity we mean validity of our inferences which is slightly different from I believe the way validity has been presented to you in previous courses – related but slightly different. I think this section, if there's no other topical set that you go back to over and over and over again, this is the one that I think you need to. It's critical that you understand threats to validity in order to allow you to make better inferences and ultimately design better studies.

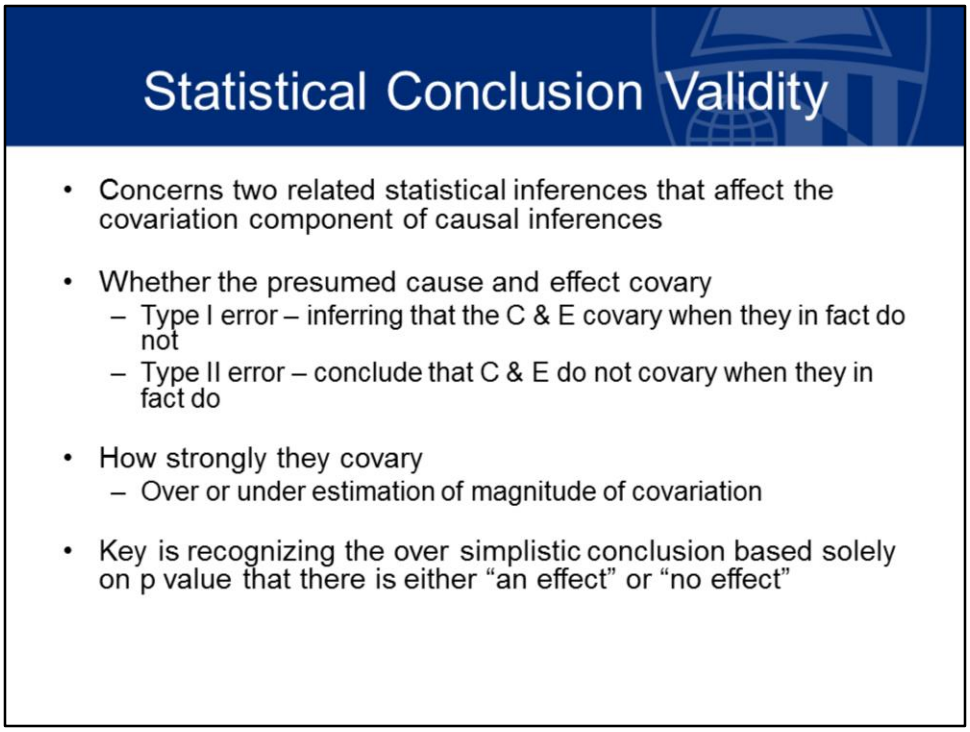


So providing a basic definition when we speak about validity here, the approximate truth or inference of a knowledge claim… and truth is it's one of those words that is difficult to nail down; therefore it's the approximate truth so the judgments we make about the extent to which the to which the evidence supports the conclusion we're drawing and whether that is a true or correct inference to make. It's important that we understand that judgments of validity are never absolute; that's why researchers have a way of caveating and being very careful with their language, especially when we're talking about causality because as in the previous presentation causality is a high bar. So it's never absolute; this is valid, this is invalid. There may be cases where we get really close to either of those two extremes but understanding that it's potentially a range. And it's also important to understand that validity is a property of the inferences we make and not of the design or methods. Virtually all designs or methods have potential to lead to valid inferences but it's really the inference itself in which we're talking about validity.



When we speak about a threat to validity we're talking about very specific reasons that we could be wrong with respect to our inference, about whether two things co-vary, whether there's a causal relationship, whether the constructs are operationalized appropriately and whether or not we can generalize to other settings or persons. So these are very specific reasons that we could be wrong in our inference, and knowing them helps us anticipate criticisms of our inferences so that we can rule them out before they arise. Going back to those plausible rival hypotheses or alternative explanations, knowing the threats to validity helps us identify which of those rival explanations might be out there and then we can rule them out on the front end through our design, addition of controls to our models, etc. But when we cannot rule them out by design controls we really have to think through how would this particular threat apply in this case? Is there evidence to say that while it's possible it's not plausible or vice versa and there's a big difference between those two of virtually everything is possible given an infinite amount of time and iteration, but is it truly plausible? Is there a high degree of evidence to suggest that it is happening in the context? And then we have to think about – and we should do this anyway – does the threat operate in the same direction as the observed effect so that it could partially or totally explain our findings, so is a particular threat leading to higher scores on an achievement test and that's the same path or direction as our treatment and when we overly those two we could erroneously conclude that the treatment is causing the effect versus this unknown rival hypothesis that's operating in the same direction. So we need to keep all of that in mind; it helps us be better and

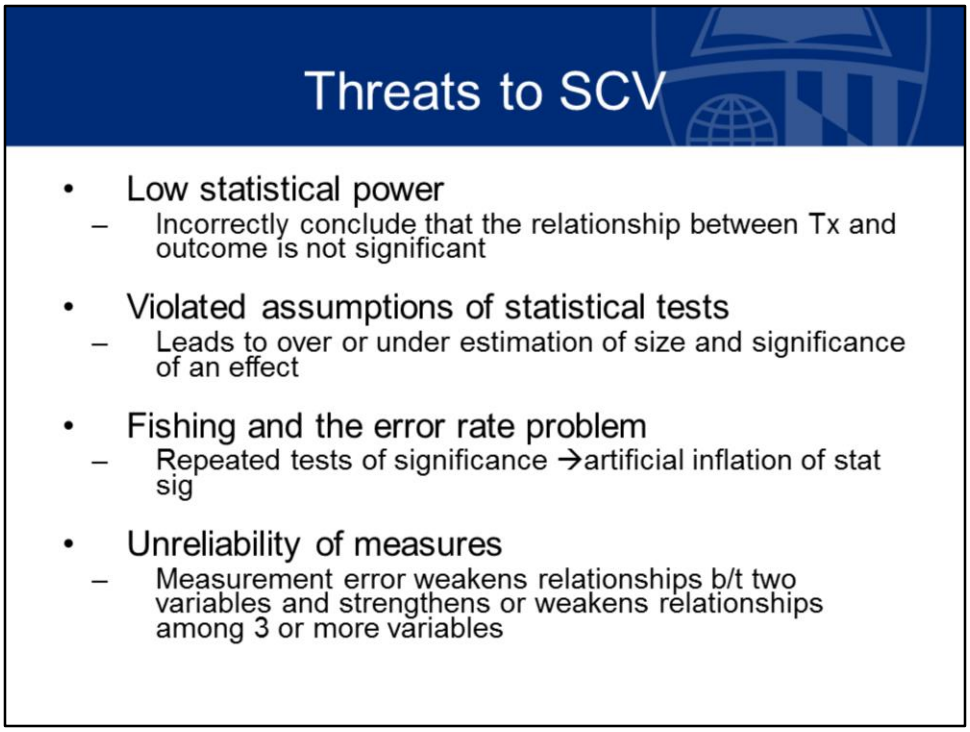
careful designers, it helps us to be better and careful reporters of our results and it helps us come to sounder conclusions about what folks should think about our findings.



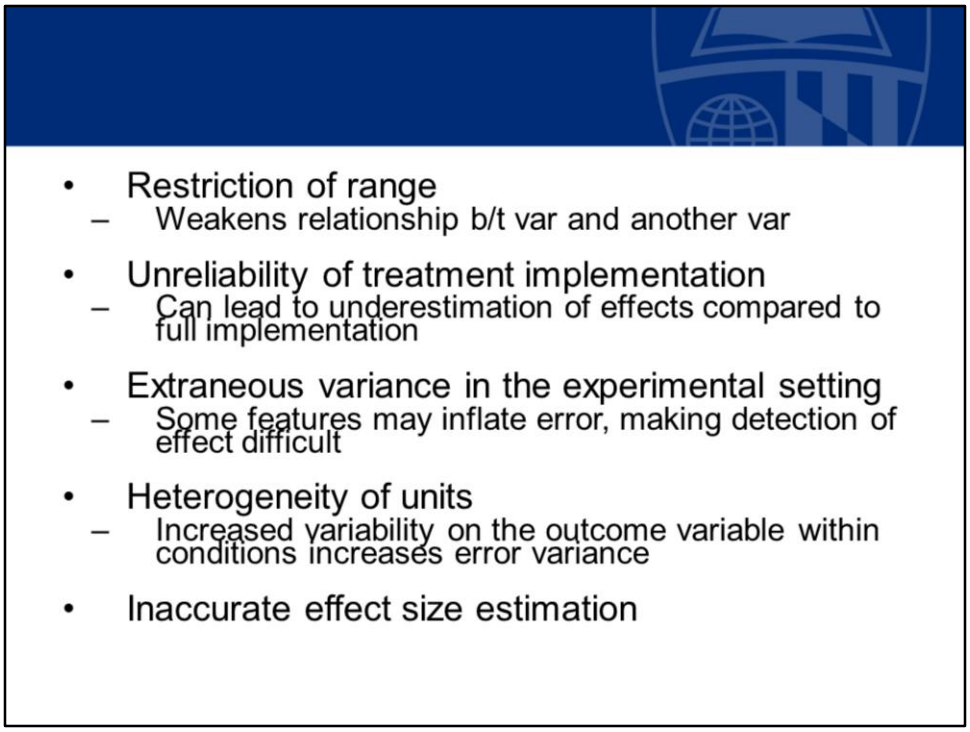
.So we’re going to now walk through the broad classes of threats to validity laid out for you in Chapters 2 and 3 in Shadish, Cook and Campbell. Again I’m by necessity hitting the high points and pedantically walking through each of the threats outlined there. I’m doing this for a reason; this is one of those topics that you should revisit multiple times, even rereading these chapters potentially multiple times. There’s a lot in them, there’s a lot to consider and it’s important that you get these threats into your head as you’re thinking about design and evaluation.

So our first section is statistical conclusion validity. These are concerns related to statistical inferences that effect the co-variation component of causal inferences. One is whether the cause and effect actually do co-vary, so revisiting your statistics knowledge is type I and type II error, so type I inferring that the cause and effect co-vary when they in fact do not, inferring that the treatment leads to the effect we see when in fact it does not; type II error concluding that they do not co-vary when in fact they do, so the failing to find a significant effect of treatment when in fact there is one. And then also how strongly this co-variation is; we could over or under estimate the magnitude of co-variation due to some of these threats.

A key point that could be said in multiple places but I’ll say it here is not falling into the simplistic conclusion that a p value alone tells you that there is an effect or no effect; it does not do that, and there’s many reasons why we could get a significant p value but there not be an effect or vice versa.



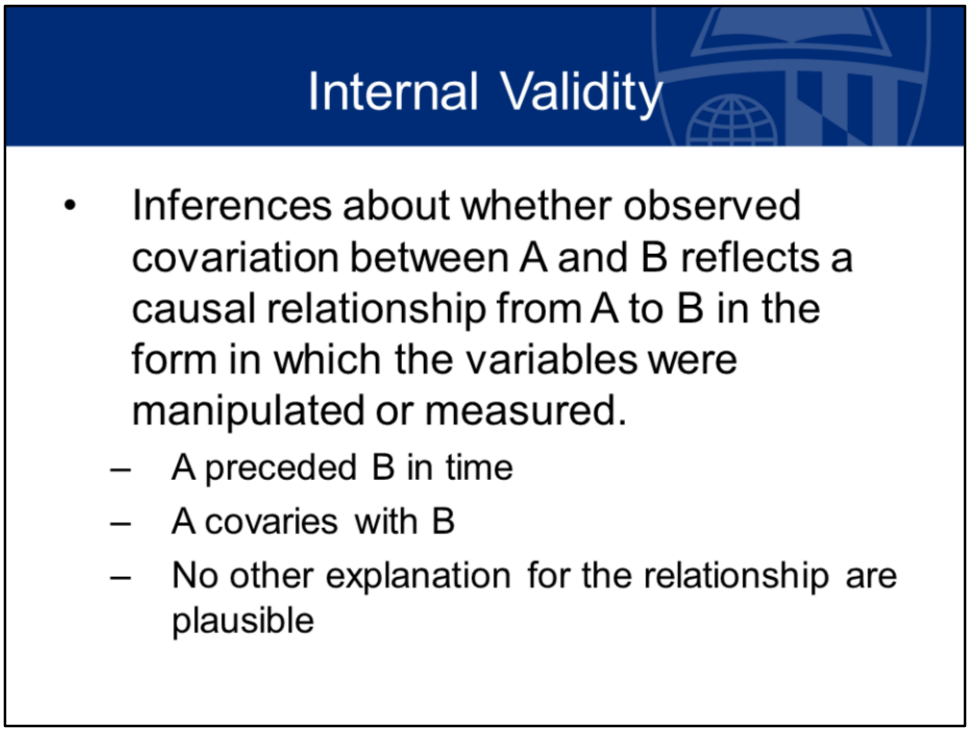
The first threat to statistical conclusion validity is low statistical power; we're going to take up power in one of the next presentations but essentially this leads us to an incorrect conclusion that there is a non-significant relationship between the treatment and the outcome when in fact we do not have the ability to detect it if it's indeed there. So we'll take that up in another presentation. If we violate the assumptions of our statistical tests, so assumptions about normality of the underlying distribution can lead us to over- under- estimation of the size and significance of an effect, fishing in the error rate problem, so repeated tests of significance known as fishing… we're going in and we're looking, trying to hunt out one single significant finding can artificially inflate statistical significance because we're not correctly accounting or adjusting for it in multiple tests. We have to consider unreliability of our measures; to the extent that a measure is unreliable this measurement error can weaken our ability to detect the relationship between two variables. If we have error in measures among multiple variables we're not particularly sure whether it strengthens or weakens the relationships among them all, so that's why in a previous course we spent a lot of time talking about validity and reliability of your measures and measurement error.



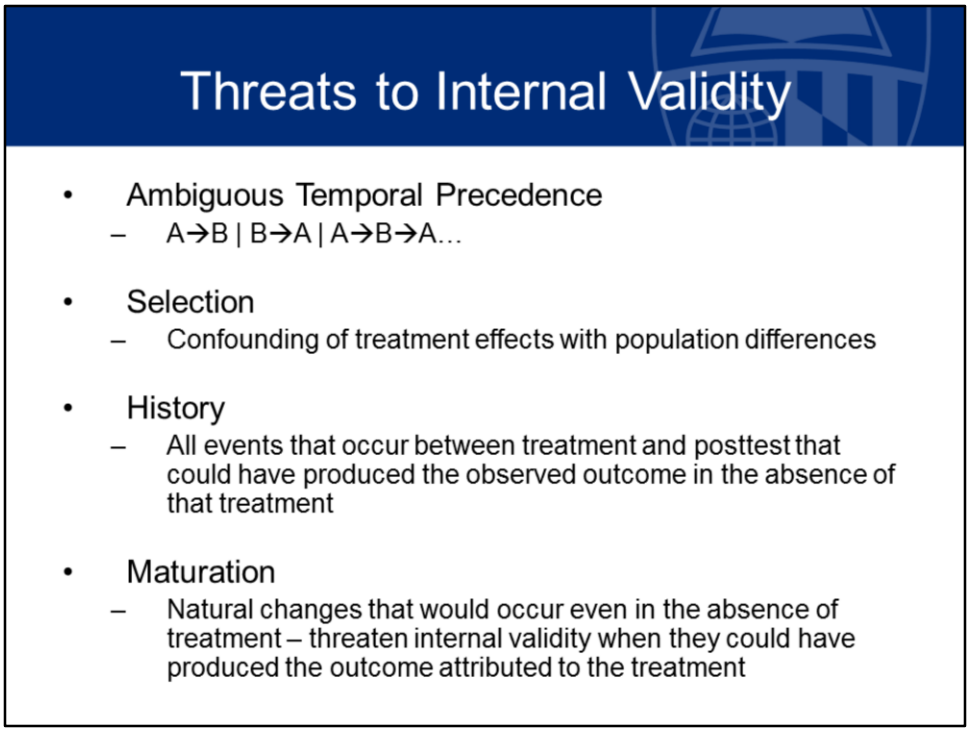
Restriction of range can lead to statistical conclusion validity threats by weakening the relationship between variables and if you think about it our variables have a range in which they vary from a minimum to a maximum and if we were to cut out a part of that range and only investigate the effect of our treatment in that range we may come to an erroneous conclusion than if we looked at the entire range, so we might conclude there's no relationship between the two when in fact if we had a full range we would actually see that. You can best see that if you think of a scatter plot and a strong relationship moving up on a 45 degree diagonal; if we were to just cut out the middle portion of that it may look like a noise of dots when actually the underlying relationship is very strong.

Unreliability of treatment implementation is a threat. As we took up previously if you are not implementing with fidelity that can lead to an underestimation of the effect compared to the full implementation so your watering down in essence through unreliability of implementation. We need to think about extraneous variance in the experimental setting, so there are some features that might inflate error making it difficult to detect an effect, so that's in thinking about how we set up our experiments or our designs, excluding extraneous variance, anything about heterogeneity or units, so differences among participants so if they differ on the outcome variable highly that could lead to increased error variance and inaccurate effect size estimation can also lead to statistical conclusion invalidity. So that's a very brief overview; please go back to the chapter, I believe it's Chapter 2, revisit those and certainly as you move forward and continue designing you'll want to think about

it.



Now we're going to talk about internal validity and threats to internal validity. This is squarely in our discussion of causality and causal relationships, so repeating again our three criterion that we need to think about… So internal validity deals with inferences about whether an observed co-variation between A and B effects the causal relationship from A to B and so we have to meet these criteria and A precedes B in time; A covaries with B and no other explanations for the relationship are plausible and that's the key one that we're going to focus in here on talking about the rest of validity.



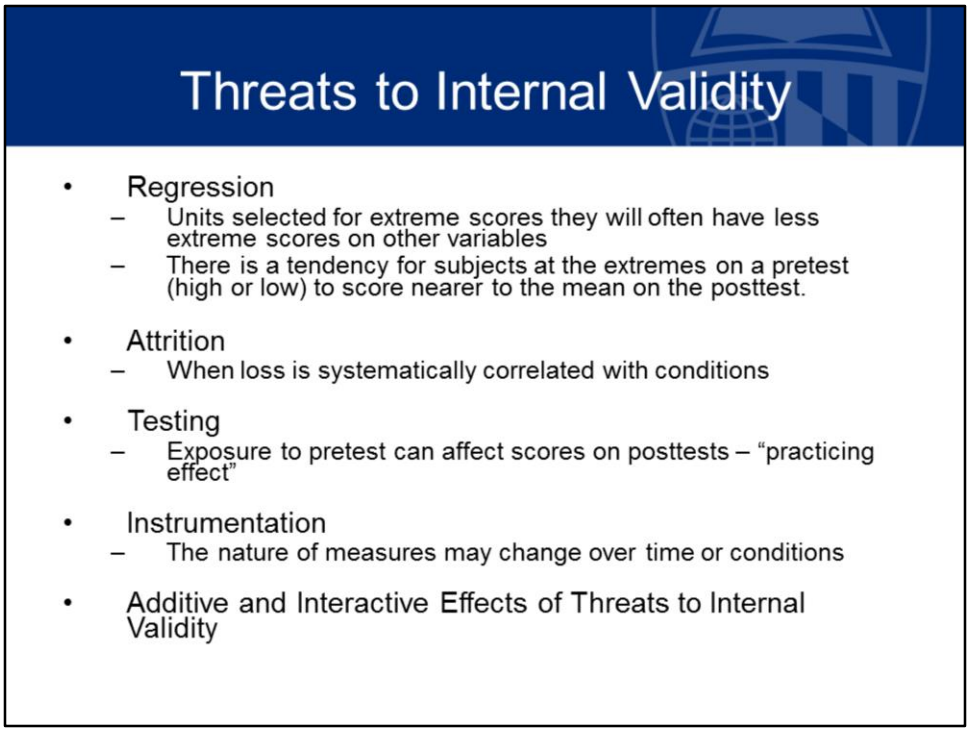
So our first threat is ambiguous temporal precedence, so it's not clear whether A precedes B or B precedes A or there's some cyclical action of A leads to B and B leads to A, so a cycle. In that case we have a clear threat to our internal validity because we can't cleanly determine that A precedes B in time.

The next one is especially important in social science. We need to think about selection; how do folks get into treatment groups and to what extent is that unobserved characteristic of them that leads them to get treatment and bounded with any outcome effects we might see. So I've given this example before; in my own work we had a potential selection threat of who chooses to enroll in a charter school in the extent that the unobserved mechanism operating within a family that pushes them to enroll in a charter school is related to higher outcomes or lower. If we just naively compare charter school students to traditional public school students we may erroneously conclude that charters lead to higher or lower academic achievement when in fact it's this selection effect that's operating and leading to this faulty inference.

The next is history, a history effect, a threat to validity by history. We have to consider all the events that occur between the time of treatment and the posttest that could have produced the observed outcome in the absence of that treatment. We need to think of history events as systematic… that it happens to all folks in the group at the same time. So for example we're providing treatment to kids and there's some traumatic event that happens to all the kids in the classroom, or on the day of the test there's a loud boom and that throws all the kids off; so it's something that

happens to all the kids that could have induced the observed outcome; in the absence of that treatment therefore we can't disentangle the treatment from this history effect. This is especially prescient in looking at if we only have one group we need to be aware of history effects; we need to look across groups if we have two groups, treatment and comparison group, and look for or guard against history effects being in one group and not the other.

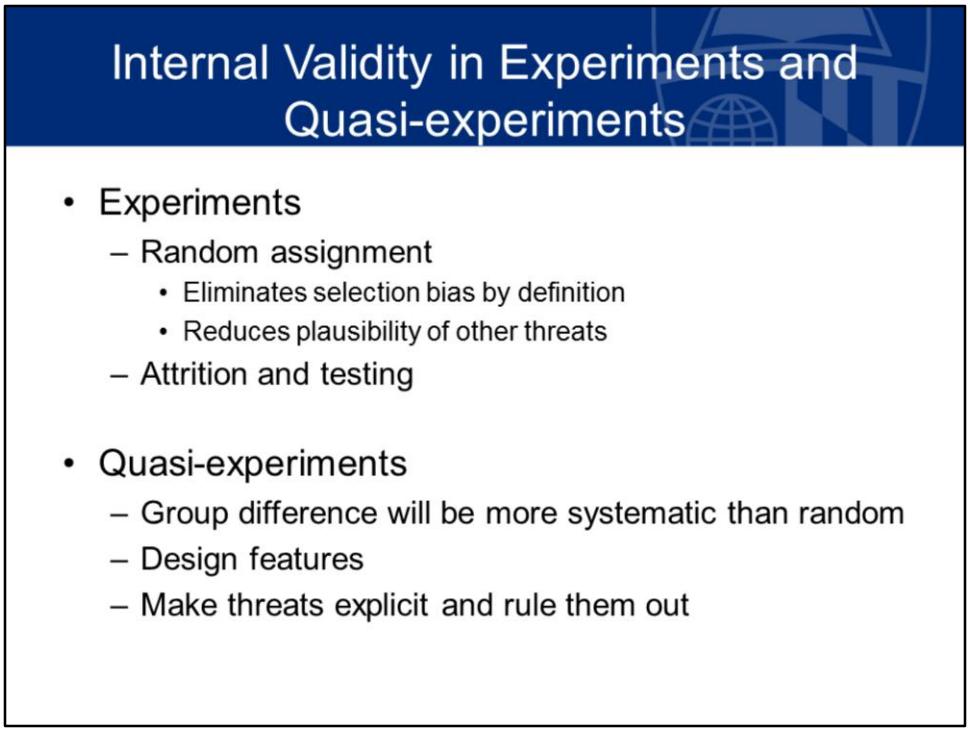
Maturation or natural changes that would occur in the absence of treatment; it's very similar to a history effect except that especially with children we can think about them developing along some trajectory that if we did absolutely nothing to them in terms of our treatment we might see a natural increase in an outcome, so we need to thoughtful about how maturation and progression are development of our outcome – actually precedes – that comes out of theory – and making sure that that is not operating and leading us to erroneously conclude that our treatment had an effect. And we guard against this generally with a comparison group of kids that are the same age; that helps us to rule out maturation threats.



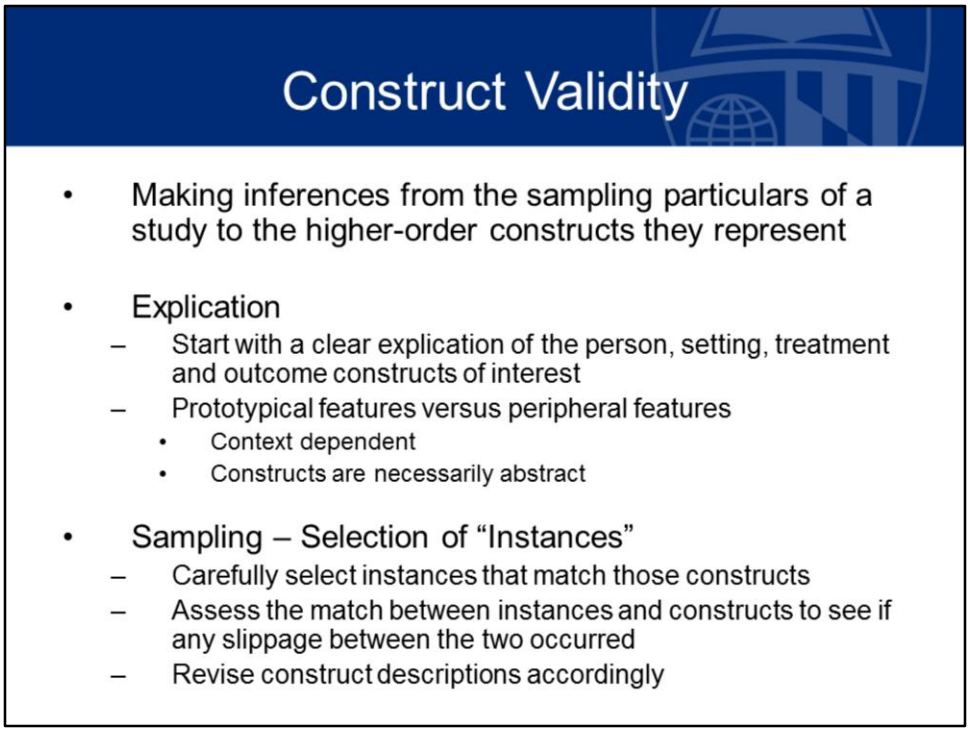
Regression or regression to the mean happens often when we select units or persons to participate in our interventions because of extreme scores; they will often have less extreme scores on other variables and if we think about - let's stick with the achievement testing outcomes – often kids will test low in one administration for whatever reason, some random noise of the day and they didn't have their orange juice or they're having a particularly bad day… that's not their true achievement; it's depressed for some reason on that given day. If we go back and retest them they're going to naturally come back closer to their true achievement level and that's regression to the mean. So we need to be careful about how we select units, whether we're selecting them for being particularly high or being particularly low, especially with the one on the test. So I gave an example for low but high could be the same; everything clicked for me that morning and I took the test and I did extremely well, above my true ability, and if I take the test again I'll come back down to my true score.

The next threat is attrition and attrition is when folks leave our treatments or leave our programs before the end of them, and we need to be particularly attentive to it – we always need to be attentive to it, but we need to be particularly attentive to it when there's a differential attrition across groups. So an example would be in that it's tied systematically to some characteristic of folks who leave to their outcomes. So for example if we have two groups, one receiving treatment, one comparison of children that are similar and we're doing some reading intervention, if we have attrition in the treatment group that is related to original pretest score so our low scorers tend to

drop out of the study, they leave, they have higher levels of mobility so we systematically lose the low performers out of our treatment group versus our control group. What we've done is that attrition has artificially inflated the posttest scores of the treatment group and that has nothing to do with the treatment, so it dropped out the bottom of the distribution and we would have a potentially faulty inference. Testing, we have to be careful that there's a practicing effect that if we give the same pre- and posttests people do remember and they could become better at taking that test and certainly we could get into a discussion about the effect of state testing and there being testing effects. There's some interesting work that suggests that when you take a test the scores go down and then as people become more adept at taking the test versus actually demonstrating the skills embodied in the test they become better at them, and so that's a testing effect and it's a threat to internal validity. Next we need to think about threats that operate through instrumentation; our measures may change over time, the way people respond to those measures may change over time and it's a function of the instrument itself so we need to think about that, and then ultimately all of these things may be operating and they may interact with one another in ways that are very difficult to predict or completely disentangle, so we need to think through each of these threats and how they might be operating in our programs and evaluations of those programs, and then think about how to design in a way that negates them before we conduct our evaluations or our implementation and evaluation.



We return to our discussion about experiments and quasi-experiments through the lens of internal validity. In our randomized experiments that randomized assignment eliminates selection bias by definition. Selection bias is something I would argue is one of the more important threats to internal validity that we need to consider when working with existing groups, so random assignment takes that out of the equation by definition. It dramatically reduces the plausibility of all the other threats. There are cases where we could have an unhappy random assignment or there could be history effects that systematically effect the treatment versus the control, but the random assignment reduces the plausibility of all those other threats. What it doesn't take care of is attrition and especially differential attrition and testing effects, so we still need to think about those even if we're doing our [inaudible word] experiment. Quasi-experiments, group difference even though we're going to put in place a mechanism that we hope approximates the random assignment. Hope only gets us so far; hopefully we do it well, but the group differences will still be more systematic than random. We need to really consider our design features here and what it really pushes us to do is be much more explicit about those threats that could be operating in our setting with the participants we have at the time, that we've done our limitation and ruled them out either through design, logic, theory, through collection of data that can help us rule them out.

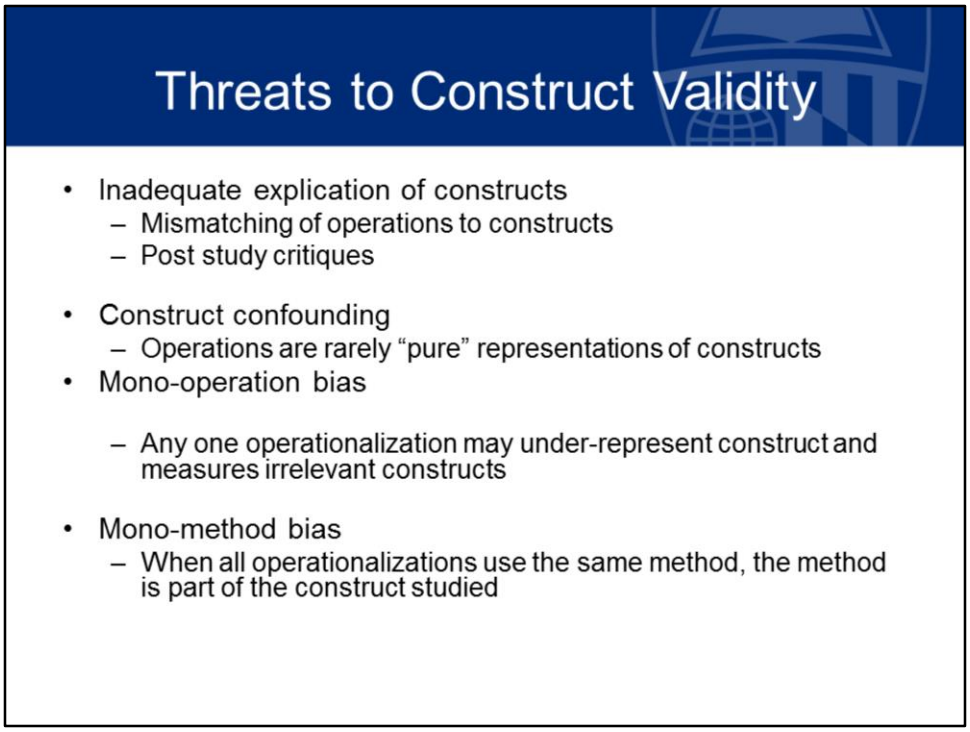


The next grouping of threats to validity we need to consider are those that can be defined in the class of construct validity and this is really where we're concerned about making inferences from the very specific elements of our study to the higher-order constructs that they are meant to represent and that we are ultimately interested in speaking to with our studies. So the sampling particulars can be how we operationalize and the very specific instruments, the folks that we hope to enroll in our programs how well did they match the constructs that we're interested in examining. So we need to start with a very clear explication of all of these particulars, the persons, the setting, the treatment, the outcome, all of the constructs represented by those elements, and we need to be very, very clear about what those constructs are and what they look like and how they might be operationalized. That's a lot of the work that goes into careful and well-designed logic models, thinking about designing our interventions. We need to pay particular attention to the constructs that we're hoping the particulars represent.

Another aspect here is also thinking about what are the core prototypical features versus peripheral features, what elements are the core no matter what context, we need to think about context dependency and realizing that constructs are abstract and we're working to make them concrete in the realized version of our studies. So we need to really have our hands around what are the core features versus those that are more peripheral and that will help us make tighter designs, it will help us be able to speak more definitively about what we're actually finding.

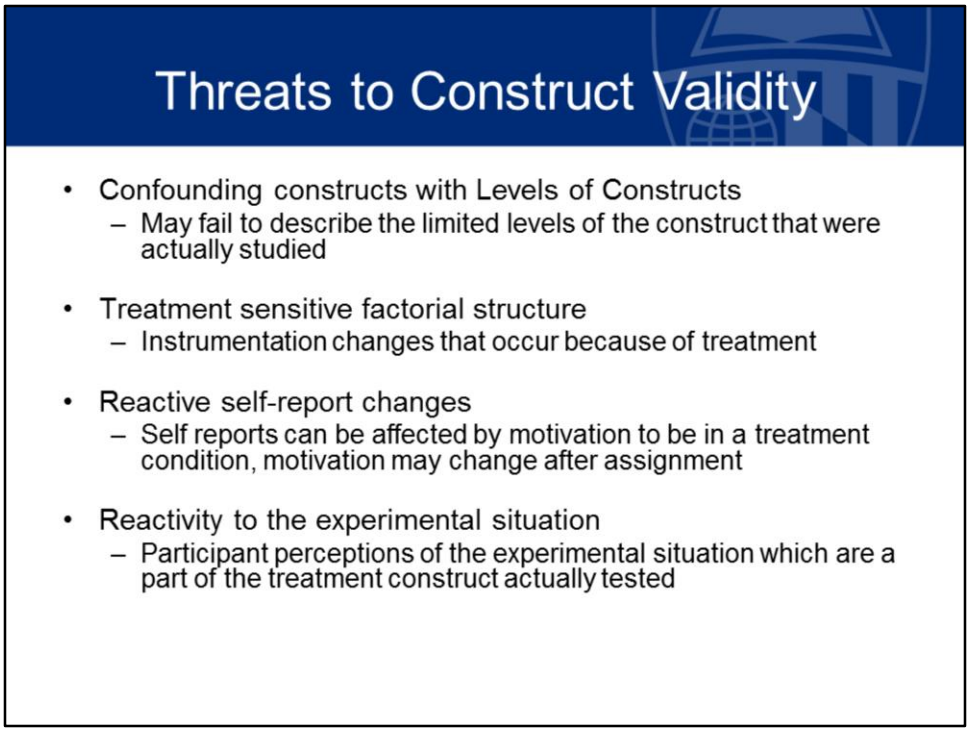
Then we need to pay attention to the selection to quote unquote instances. And so

starting from a strong base of understanding the constructs and the interplay between them we can be very thoughtful in selecting the instances that match those constructs, be that the very specific operationalization of certain constructs, how we might go about measuring them. This puts us in a place to really think about how well our instances match our constructs and whether or not there's slippage between the two so that we need to be careful that we don't… If we think we're examining construct X but due to the particulars of building our studies and enrolling folks we end up actually having construct Y and we don't recognize the slippage that has occurred between those two it may lead us to erroneous inferences, and really we need to pay attention throughout and revise where needed.

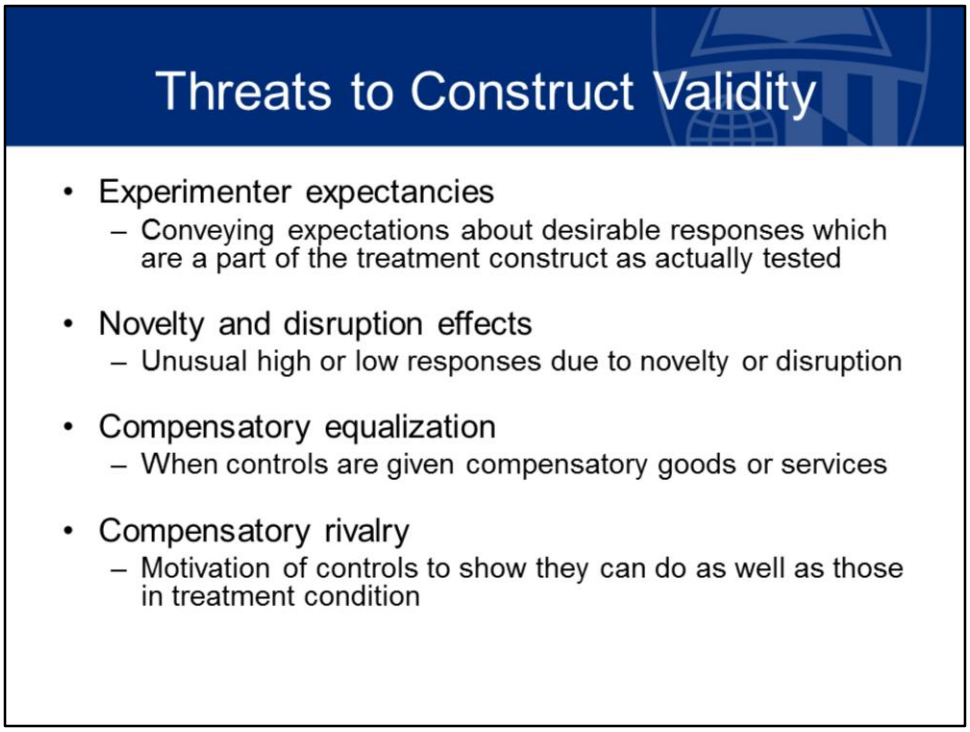


So specific threats to construct validity; one is when we've inadequately explicated our constructs; we've made an mismatch between our operationalization so the actual constructs we're intending such that we might when we collect our data we're actually examining construct Y instead of construct X. We need to be careful and think through the underlying theory of treatment, that's why we spent time on that. Another threat is construct confounding. Constructs are abstractions and to the extent the quality of our refining that to an actual operationalization, we need to be careful and paying attention to have we actually pulled out multiple constructs or does our measure actually… is it measuring multiple constructs because we haven't fully refined and done the actual operationalization well.

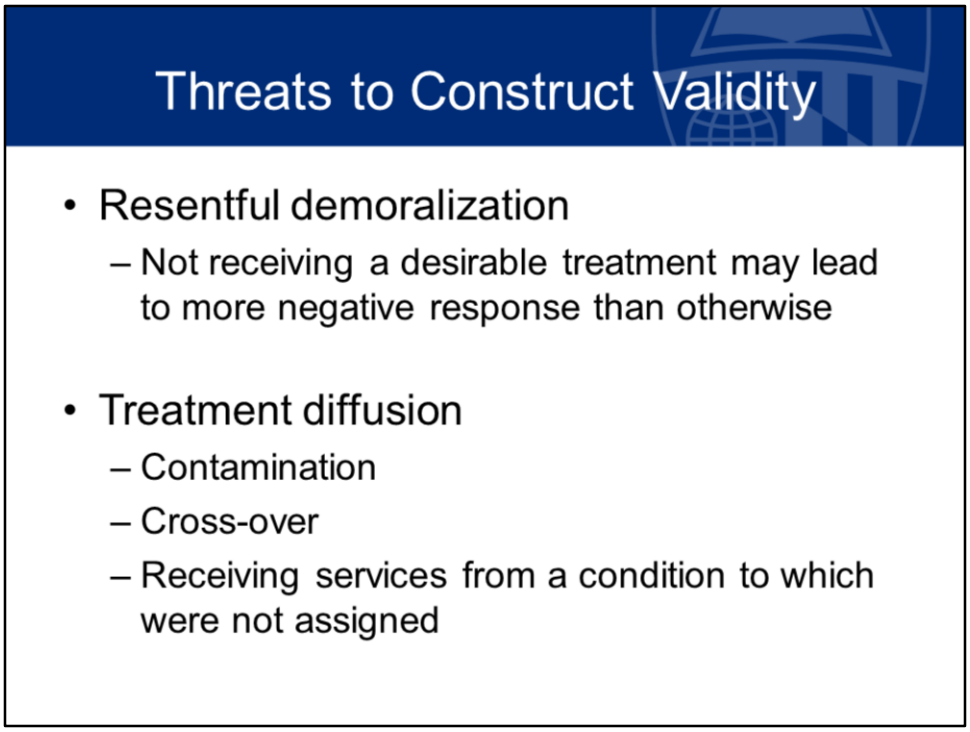
We also need to be careful about mono-operation and mon-method bias. So mono-operation… let's take reading as an example. As a meta-construct reading you may not have particular substantive pieces so we have to break that out into different sub skills perhaps and if we choose only one operationalization of the construct of reading skill it may underrepresent the entire construct of reading. Similarly if we only gather our data through one method we can also introduce bias that perhaps our answer is in part generated by the underlying construct but maybe there is an influence of the method itself, so thinking about do we collect data through observations which potentially brings in a certain set of issues, do we do it from surveys; that may introduce another set of issues when we're asking people to self-report on things. So we need to think about those and how they impact or threaten construct validity.



In the next several threats to construct validity we need to consider ways that the study itself can change or threaten the validity of constructs. One could be treatment sensitive factorial structure, and this is where treatment itself may change the nature of the construct or the way that the construct has been measured. Treatment may differentially impact separate components of the constructs but if we're only creating a measure of the original construct we may miss these changes in structure within itself. We need to think about reactive self-report changes; people may have a differential reaction to being placed in a treatment condition and this may impact the way we self-report, and we also need to think about other types of reactivity to the experimental situation so by telling somebody they are part of a treatment they may exhibit some placebo-like effects of just telling them that alone may induce better performance and we have to realize that's a part of the treatment construct as well, potentially, and it's not just some pure realization of the treatment construct.



We also need to be wary about ourselves introducing threats to validity through our own expectancies about what should happen. We may inadvertently convey expectations about desirable responses to participants and this may induce changes in participants over and above whatever the treatment effect may or may not be, so that's a threat to the validity of the construct itself, so treatment includes part of this as well. We need to think about novelty and disruption effects. Often we're introducing our interventions in ways that disrupt normal business as usual and this may be a part of the treatment construct. We need to think about compensatory equalization, so sometimes we may have situations where well-meaning but perhaps misguided thought of giving compensatory goods or services to folks in the control condition to make up for the fact that they lost the random assignment to treatment and this changes the nature of the constructs we're investigating. We need to pay attention… and I think we touched on it in fidelity… of just calling the comparison business as usual and not understanding what goes on there is potentially problematic in that we really need to understand the control condition, the comparison condition and what is actually happening there so that we can investigate whether there are compensatory equalization things happening.
Similarly we also have compensatory rivalry that may be operation so some folks by being placed in the control actually work harder than they would have in the absence of the study and this changes the nature of the construct. I've seen this often of folks saying, "Well we'll show them; we'll do better than the treatment," and this becomes part of the treatment itself and we need to account for it and think it through.

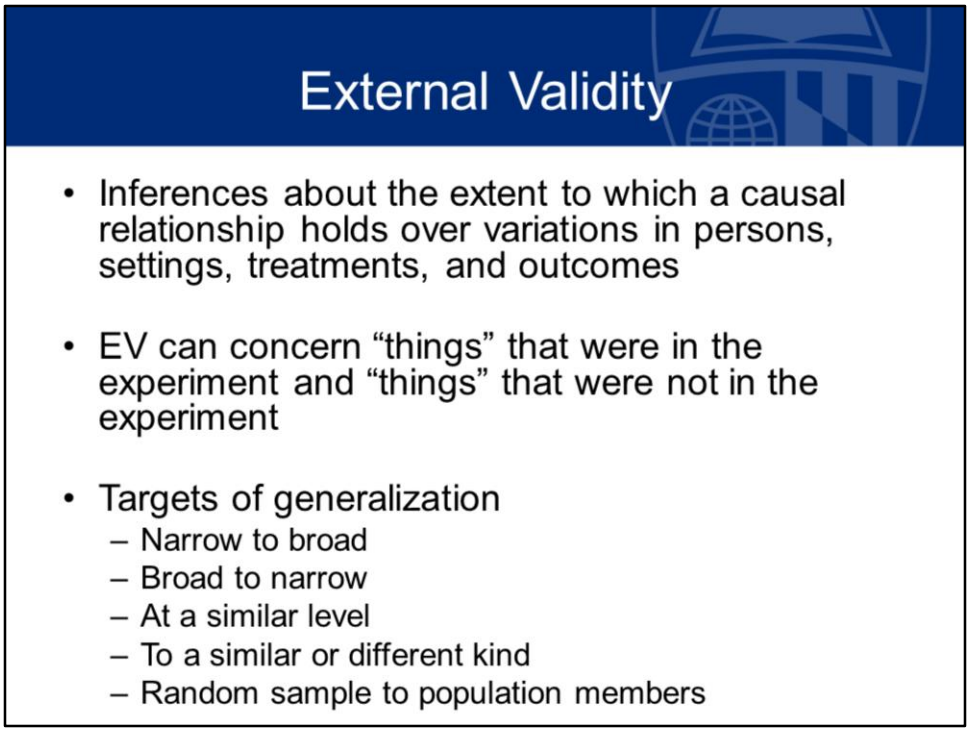


Similar to compensatory rivalry where the comparison group works hard and performs better than they would have in the absence of the study we can also get resentful demoralization; we're not getting the treatment condition that one wanted, perhaps, leads to a more negative response than what would have happened in the absence of the study, so we need to think that through as well and come up with ways of observing it if it's there.

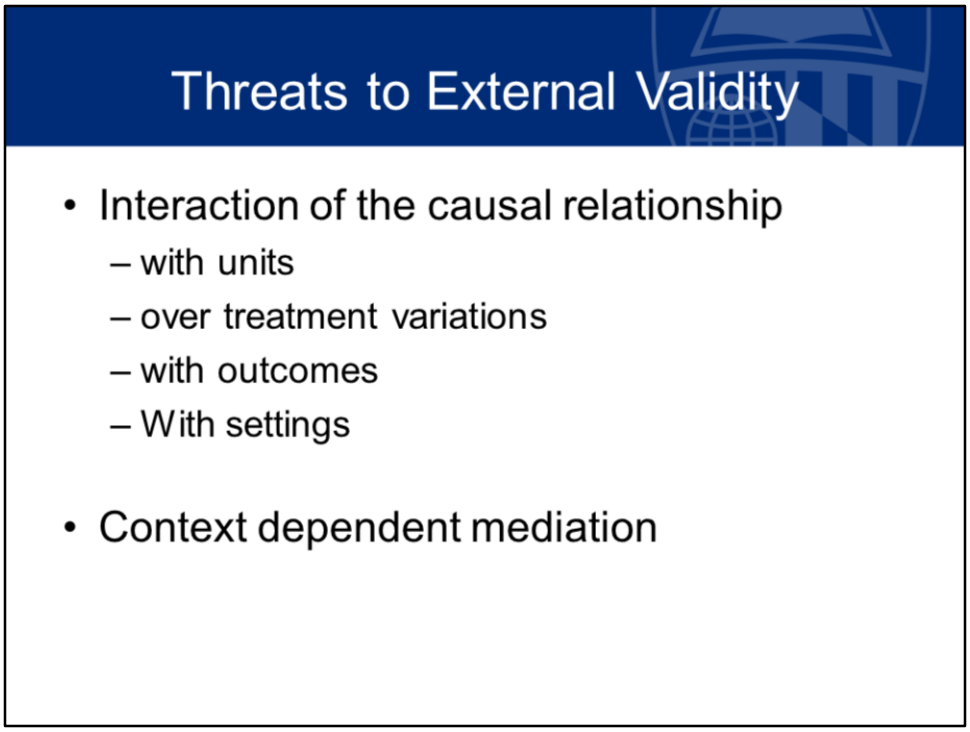
And finally treatment diffusion; well-meaning folks often don't do what we would like them to do, especially if we're thinking through providing an intervention to one group and not to another so we need to think about ways that the treatment can actually show up in the comparison group. So contamination is one form of treatment diffusion where elements of the treatment or intervention end up appearing in the comparison group. I think I may have mentioned before about this reading program, early grades reading, where kids are matched based on their tested reading abilities; they'd match a high child with a low child and they worked together to assist each other to do better. This group that created this intervention had been doing it for many, many years in Nashville and elements of that treatment became well-known in the district and actually contaminated any potential comparisons as teachers that kind of got loose in the wild and teachers tended to do it so there was no true non-intervention for comparison.

Similarly crossovers; we need to watch for people that change treatment status outside of the experimenter or researcher moving them, so folks that are assigned to a comparison group but find a way to receive the treatment itself, either through

political wrangling or somehow finding the services. And so all of that is we need to think through how we observe both conditions and do we have the tools available to us to observe pieces of the treatment in the comparison group.

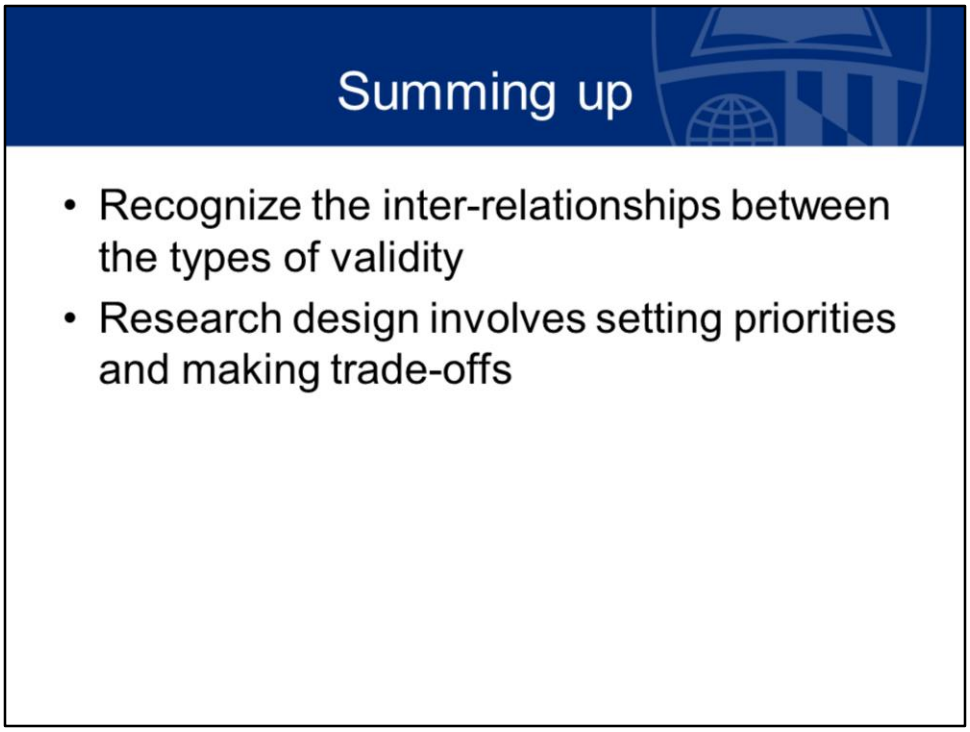


External validity is concerned with the extent to which an inference about a causal relationship that we find in a particular study or context – the extent to which that holds across variations in different types of people, perhaps settings, different treatments that are targeting the same construct and the other outcomes. So this is often called generalization so this is can we take the findings from this particular study and move it to a different context and make the same inference that it will work there. So we need to remember that external validity is concerned about the things that were in the experiment itself, so the particulars of that study, and also the things that were not in the experiment and so thinking through those two things and then how does that then move over horizontally to a different setting at the same level or vertically say from a particular instance of a population of students in a district to all U.S. K-12 education, that kind of thing. So we need to realize that generalization has various targets; we can go from the narrow instance of the study to a broader generalization; we can also go the reverse. So we need to think through what is our target generalization and to what extent does it affect whole… from the particulars of a certain study to this target.



Threats to external validity often come in the form of an interaction between the causal relationship that was found and the particulars of the study itself, so an interaction with the units, the people that made up the participants, perhaps the schools, the districts, whatever, interactions with treatment variations with the outcomes and with the settings. So all of these particulars could potentially be driving some of the causal relationship; to the extent that there are some interactions there it may lead to a lessened ability to generalize to other units, to other variations of treatment, outcomes and settings, and so we need to think that through. An example perhaps is often in charter school research that's concerned about academic achievement outcomes we have to utilize the fact that when oversubscribed schools have to have a random lottery to assign seats to kids so some kids get seats and some kids don't; it's a random process which helps increase our internal validity of the study because the only mechanism there is the random assignment. But we have to be careful because what we now have is we've reified a bit to only studying schools that are over-subscribed; if they're not over-subscribed there's no random lottery. So there's a different kind of unit, so only schools that are doing something well enough to be over-subscribed are the focus of the study and then we have to ask well does any finding of a positive charter effect in that situation actually apply to all charter schools even if they're not over-subscribed. So we need to be careful about that and think about all of these things as we're hoping to make a generalization.
And finally we need to think about context-dependent mediation, so perhaps there's a mediator that is operative in our particular context but in another context it may

not be operative. So all of these things are tied together and ultimately they come back to how well have we understood the theory behind our treatments and the ways that might guide us to think about variations and how they may or may not apply in different settings, etc., etc., so we need to think that through and it all comes back to that base.



Summing up it is important I think that you revisit these chapters in the Shadish, Cook and Campbell book, return to them often, read and reread sections, think about how these different types of threats to validity apply to your particular studies. They're crucial; you need to get your hands around them, you need to understand them, you need to understand them in the context of your study because all of the things you're hoping to do rest on understanding these and your ability to provide strong answers to the questions you're asking really lays on this foundation of understanding these threats to validity.

So again I'm going to say it again because it's important… I want you to read and reread these chapters and read them well. Another thing we need to realize is although we've talked about these discretely there's interrelationships between different types of validity: internal, external, and that our research in the design work, we need to set priorities. Are we after testing a causal hypothesis in the strongest possible way so that at the end of the study we can from a position of strength, of strong internal validity, make an inference that X causes Y? That will involve tradeoffs in the research design and that priority setting will drive a lot of the way that we create the study, so we're going to be going for a randomized control trial and we're going to pay attention to those issues.

Conversely perhaps we're most interested in generalization and that priority setting might lead us down a different path, but it's important to realize that we must have a priority; we must set priorities and this involves making tradeoffs and we need to have thought through, again, all of these things, all these issues in order to set those

priorities and then make appropriate tradeoffs.
And then I would also point out to you that somewhere around page 97 in the Shadish, Cook and Campbell book there's a nice discussion about the interplay between internal validity and external validity and how we think about the tradeoffs that might be necessary. We certainly want to pay attention to internal validity so that we come to the strongest possibility for our inferences about the effects of our treatments, but we must also recognize that generalization and external validity is also very important and we need to think those issues through.

Ok, I'm going to say it one more time to make you sick, but read these chapters.