INSTITUTE OF EDUCATION SCIENCES

FIFTH ANNUAL IES RESEARCH CONFERENCE

CONNECTING RESEARCH, POLICY AND PRACTICE

CONCURRENT PANEL SESSION IV

"Generating Plausible Hypotheses Using Difference-in-Differences Estimation"

Wednesday, June 30, 2010

10:52 a.m.

The Gaylord at National Harbor
National Harbor 12-13
201 Waterfront Street
National Harbor, Maryland 20745

CONTENTS

- - -

PROCEEDINGS

DR. RUBY: Good morning. I am Allen Ruby from IES. The impetus for this session today on difference-in-differences estimation was twofold. First, we’ve noticed an increase in its use in grant applications. However, they’ve been submitted almost solely by economists. So, we wanted to raise the question whether this is a method that might be useful for researchers in other fields.

Second, experiments are expensive, and so, we were wondering, could difference-in-differences estimation with its use of secondary data provide additional evidence regarding the impacts of an education intervention, and so, could it provide at a lower cost way a means to determine whether an experiment should be done, and, even under certain conditions, could it replace an experiment?

The goal of this session was to explicate difference-in-differences for researchers for many fields and also to evaluate the evidence that can be attained from it, and, to this end, we asked Larry Hedges to present on it.

Larry joined the Northwestern faculty in 2005 and is one of its eight Board of Trustees Professors at Northwestern. He has appointments in statistics and education and social policy. He was previously the Stella Rowley Professor at the University of Chicago, and his research has involved many fields, including sociology, psychology, and educational policy, and he’s especially known for his work in methods for meta-analysis.

Putting in a plug for his most recent 2010 publication in the journal *Research Synthesis Methods*, in which he addresses nonindependent effect-size estimates when doing meta-analysis.

He has served in an editorial position on a number of journals including the *AERJ*, *American Journal of Sociology*, *Journal of Educational and Behavioral Statistics*, the *Review of Educational Research*, and others, and he’s been elected a member or fellow of numerous boards, associations, and professional organizations, including the National Academy of Education, the American

Statistical Association, the American Psychological Association, and the Society of Multivariate Experimental Psychology, and, most recently, he was elected president of the Society for Research on Educational Effectiveness.

In response to our request for this session, Dr. Hedges has reached into his expertise in both design and analysis to consider the quality of evidence provided on the assumptions used with quasi-experiments, including difference-in-differences estimation.

Please welcome Dr. Larry Hedges.

[Applause.]

DR. HEDGES: Thank you, Allen, and especially thank you for the plug for my most recent research. It’s always nice to get that. When I was asked to talk about difference-in-differences, well, actually they asked me a long time ago, I said “yes” because it was a long time between then and when I would actually have to do anything.

[Laughter.]

DR. HEDGES: And after thinking a little bit about what I might say—I realized that it was, I thought it was important to say—to talk about something more than just difference-in-differences. So, to give you a sense of what I had hoped to speak about today, I made a slide of goals.

What I intend to do—because, I was asked to talk about difference-in-differences and setting them in the context of plausible causal inference, possible causal inference—I realized I really couldn’t do that without talking a little bit about the modern theory of causal inference.

I’d like to begin by speaking with you a little bit about that. Then, I’d like to make the obligatory explanation in that framework of why experiments provide model-free estimates of causal effects, and I’m going to emphasize the model-free aspect of inference for causal effects because there are lots of ways to get model-dependent estimates on causal effects—of causal effects.

The problem is model-dependent estimates of causal effects are only estimating the right thing if the model is right, and, frequently, it’s difficult or impossible to know if the model is right.

I want to turn from the easy case of experiments to the much more difficult and model-based, in most cases, situations of quasi-experimental designs. I’m going to talk about three because they kind of lead up to difference-in-differences in an important way.

One is assignment based on a covariate. Another is the regression discontinuity design. The third is the nonequivalent control group designs in which difference-in-differences kind of sits in a certain way, and then spend the remainder of the talk looking at the difference-in-differences approach in greater detail.

With that in mind, I think it’s important to all of us at this conference to think about causal effects. I mean, we all, I think, claim that we’re interested in finding out about the actual causal effects of the interventions and other objects of our study.

Everybody kind of thinks they know what they mean by cause and effect and causal effects, but a formal treatment is actually useful, I think, and it actually helps to analyze situations in which the basis for causal claims is somewhat more tenuous than it is in randomized experiments.

The modern approach to causal inference is sometimes called the Rubin model or the Rubin-Holland model or the Rubin-Holland-Rosenbaum model of causal inference. It’s important, I think, to recognize, though, that these ideas are very fundamental, and people had been thinking about them well before Don formalized them in the ’70s.

In fact, it’s interesting that the roots of modern way of thinking about cause and effect go back to that fertile period in the 1920s when statistics was, as we know it today, was really being developed.

I think everybody knows that one of the great triumphs of 20th century statistics was the development of the randomized experiment and the development of an understanding of why the randomized experiment gave unbiased estimates of causal effects.

Most people don’t know that Jerzy Neyman gave a very modern account of why agricultural experiments should yield unbiased estimates of causal effects. Fisher had his own account, and that’s the one most people know. What most people don’t know is that Jerzy Neyman, who at that point was a young Polish statistician, gave a remarkably modern account of how to think about causal inference in a paper in 1923 that was largely ignored until very recently.

Now, for those of you—as a minor digression—for those of you who don’t know the name Neyman—you can’t quite place it in the pantheon that you can place Fisher in—Jerzy Neyman was a remarkable statistician, a remarkable scholar, somebody whose stature is—he’s one of the few people whose stature you could say is really equal to R.A. Fisher, and his accomplishments are really equal to R.A. Fisher’s.

You may be—it’s interesting that he made contributions to experimental design in thinking about causal thinking, but Neyman is also associated with the other great triumph of 20th century statistics, and that is the development of the theory of probability sampling and the account that all of us are now familiar with, of how probability sampling allows you to make model-free generalizations to populations, and there’s an important parallel here.

Experiments allow us to make—randomized experiments allow us to make model-free estimates of causal effects. Probability sampling allows us to make model-free generalizations. And it was, that is, so commonplace now that probably all of you think that statisticians always accepted surveys and probability sampling idea as just absolutely natural.

Well, in the beginning of the 20th century, statisticians didn’t think of surveys as natural. They didn’t think of sampling as a good way to get information about populations. And it was only after Neyman gave a very detailed mathematical account of how this whole inference process could work that surveys became accepted in the statistical world.

So, Neyman has a key role as one of the people who is associated with the other triumph of 20th century statistics besides experimentation, and Neyman was a mathematical genius of the same order as R.A. Fisher. You might be amused to know that the same paper in which Neyman gave the fundamental basis for probability sampling and generalization from probability sampling is also the paper where confidence intervals were invented.

All of you are familiar with that. You might even think Fisher invented it, but, in fact, Neyman invented it, and it was a side thought in this major paper on surveys. And just to say two other good things about Neyman, one is that in the early 1930s, Neyman was one of the key developers of many of the fundamental aspects of hypothesis testing that all of us take for granted.

For example, although Fisher advocated null hypothesis testing, he was not the guy who invented the concept of power or Type I and Type II errors. It was Neyman, along with Egon Pearson, who developed the theory of the decision-theoretic basis for hypothesis testing. The idea of power is a Neyman idea, not a Fisherian idea, and as a matter of fact, they not only developed that concept but discovered the fundamental theorem on ways to find most powerful tests.

Neyman is a remarkable guy as a scholar, and it’s interesting that his work intersects this early work—was, actually anticipated this early work on causal inference.

I’ll say one other thing about Neyman is that those who knew Fisher—I didn’t—said he was quite, well, they said he had a keen appreciation for his own genius.

[Laughter.]

DR. HEDGES: Neyman, on the other hand, I did know slightly as a graduate student. He was still tottering around Berkeley at the time and occasionally would rusticate in Palo Alto at Stanford, and Jerzy Neyman was really a sweet guy, you know. That wasn't just my experience. Everybody who knew him saw him as a relatively humble decent fellow who treated graduate students and staff as well as he treated the most distinguished of his colleagues, which, I think, is a testament to him as a man.

Okay. Now, I've ended my digression. Let's go on to talking about causal inference. The key concepts in the sort of modern causal inference theory, which actually goes back to Neyman, but we usually call it the Rubin-Holland model, is that there are units—I bet this thing is a pointer. Oh, no, it isn't. Is this a pointer? Yeah, that's a pointer.

The key concepts are units like individuals, treatments. I'm going to assume two treatments, but it's easy to assume more than two treatments. It just makes the world more complicated, and so, for the purposes, it makes the notation more complicated—for the purposes of our discussion, it's sufficient to have two treatments. You might call one of them the intervention group and the other one the control group.

There is another key concept: responses. And the basic idea is that each unit has two possible responses if there are two treatments. r0 is the response that unit would produce if they got the comparison treatment—the control group—and r1 is the response that unit would produce if it got the treatment.

Then, having defined these two possible responses associated with each unit, we can define the causal effect of treatment 1 versus treatment 0 on unit i, and I use the symbol Tau i to indicate the treatment effect on unit i.

Well, treatment effect is just the difference between what you would have observed if that unit had gotten treated versus what you would have observed if that unit hadn't gotten treated. That's the sort of fundamental definition of the causal effect of treatment 1 versus treatment 0 on unit i.

There are some things to notice about this definition. I've repeated it up here. One is it's a relative definition. Treatment effects are always relative to something else, and so, to say treatment effect without a counterfactual that is attached to it is an incomplete statement.

Secondly, this definition of treatment effect is a counterfactual definition. It’s counterfactual because you can’t observe both what happens to the unit when it’s treated versus what happens to the unit if it’s not treated.

There’s actually, and this leads to the conclusion that the relative causal effect of treatment on a single unit really can’t, you know, can’t be estimated without additional assumptions.

Now, there is one interesting situation in which additional assumptions can be made that seem pretty plausible, and that’s the single subject design world in which repeatedly the same unit is made to experience both treatment conditions, and, although it’s not a model-free way of estimating treatment effect on an individual unit, it is sort of a notable attempt to try to get at treatment effects on individuals, but, except for situations like that, we really don't—we really can't have the ambition of estimating causal effects of treatments on individuals.

One of the things that's important to point out here—and this is more for your general cultural enlightenment perhaps than anything else—is that causal inference is fundamentally a missing data problem.

That perspective is natural to some people but sounds bizarre to other people, but when you think about it, it's very sensible. The problem with causal inference is you can't observe both what happens to a unit if it's treated and what happens to the same unit if it isn't treated.

So, in a sense, trying to estimate Tau i—the treatment effect on the ith individual—is a problem that's easy to solve if we can observe both r0 and r1 for that unit, and it's difficult to solve if we can't observe them both, and that's exactly a missing data problem.

Consequently, it shouldn't surprise anybody that modern ideas for causal inference sometimes draw on modern ideas for handling missing data, and it’s not a coincidence that Don Rubin, who is interested in causal inference and kind of has been a big proponent of this model, also is somebody who made part of his career working on missing data problems in statistics.

Essentially, missing data methods attempt to find conditions that reduce the missing data to be conditional on a set of covariates, conditional on some observed data, as if they were random, and I don’t mean to trivialize the way these methods work because they’re actually quite sophisticated and quite useful, but when you boil down what we try to do in many modern missing data methods, we try to figure out conditions under which we can treat the data that’s missing as missing at random, usually conditional on a whole bunch of things we observe.

Well, because of the parallel between missing data methods and causal inference, modern methods for trying to do causal inference in nonexperimental situations frequently, in fact, almost always, at some level of abstraction, try to reduce the problem of causal inference to saying that essentially the missing data that you need for causal inference is in a sense missing at random or the assignment is, to put it another way—which I won't draw the connection too carefully in this talk, but, trust me, it's easy to draw—that the modern methods for handling causal inference tend to try to construct a situation in which conditionally, given what's known, the assignment to treatments is "as if" random, and we'll talk a little bit more about some of these things later.

But I want to give you an example of the quite literal understanding of these potential responses in the Rubin-Holland causal model. The basic idea is that every unit is running around with these two responses. I mean that's the sort of conceptual framework. We can't observe them all, but every unit has got them.

Here's a case in which we have eight units, and each, I've written in this table—I played the deity here or played the constructor of the world, in some sense, this little miniworld—each unit has a response under treatment and a response under control, and, because I know these two, I can compute the causal effect of the treatment on that unit.

The way I constructed my little world here, in the first four units, the causal effect of treatment is plus 10, and, in the last four units, the causal effect of treatment is minus 9, and this—I wanted to write down this data set just to illustrate a couple of things.

One is that the causal effect of treatment doesn’t have to be the same for everybody. Frequently, it isn’t. Also, if we imagine what a real experiment would be like, a real experiment would, or even a real quasi-experiment would, assign units to one or the other of the treatments, and thus, we would observe r1 for some of the units. In this case, units one, five, six, and eight are assigned to treatment, so we observe r1 for those units.

Units two, three, four, and seven are

assigned to control, so we observe the outcome under control, and for none of these units can we directly estimate the causal effect of treatment.

By the way, because I've sort of grayed out the things we don't observe there, but if you sort of look overall, you can see that the causal effect of treatment is about a quarter on the average, and you would also notice that in this particular experiment, the causal effect of treatment wouldn't be quite a quarter.

But, if we were to average all of those causal effects, that one-quarter that I talked about is the long-run expected estimated causal effect from an experiment.

I just want to make—well, I made that point there—the average overall possible treatment assignments would be our average causal effect. Sometimes, people, you know, talk about wanting to study causal effects by assigning people to the best treatment, and I just wanted to point out here that assignment to the best treatment in this little experimental world wouldn't—is not a very good way to get at the average causal effect—and you can verify, if you wish, that if we assign the units that have the best outcome under treatment, that is one to four to treatment, and the units that have the best outcome under control to control, that is units five to eight, the estimated causal effect here would be zero.

And that’s not the right answer, and it just illustrates that sometimes, something that sounds like a really good idea as a research strategy may, in fact, not be such a good idea.

To go on a little bit about randomized experiments and ask the question, “Why do randomized experiments give estimates of the average causal effect?” The usual—the usual strategy here is to define an assignment variable Z and say that we’ll treat Z as zero if the unit is assigned to control and Z as 1 if the unit gets the treatment.

Random assignment means—and this is the technical answer for why randomized experiments give you the right estimate of the average causal effect—the average value of the outcome under control for people who get control—that’s what this conditional expectation is—says it’s the average of the r0’s, the outcomes under control for the people who get the control. It’s the same as the average value of the control outcomes for people who get the treatment, and it’s equal to the average overall of the control outcomes.

Similarly, the average of the outcomes, if you get the experiment, for people who actually do get the—I’m sorry—the average of the outcomes people would have gotten under treatment for the people who get control is the same as the average value of the outcomes under treatment that the people who get treatment actually get, and that’s equal to the average value of all of the outcomes under treatment.

One of the things that—if you recently had an elementary statistics course, and you remember hearing about independence or conditional probability, these two statements imply that these potential responses are independent of assignment.

This is a probability theory statement.

One of the ways of saying what’s special about randomized experiments is that assignment is independent of the outcomes, and that means that the average—there should be a 1 on that guy—the average—the reason why randomized experiments give model-free estimates of the average causal effect is that the independence of assignment and the potential outcomes means that the difference between the average outcome—the average of the difference in outcomes between “if you got treatment” versus “if you got control” is the same thing as the difference of the averages.

The argument here is the average of the difference is the same as the difference in the averages, and each of the components of the difference in the averages actually is the same as the average for those who got treatment or the average for those who got control.

This is a little proof that randomized experiments have to give you unbiased estimates of causal effects, and it involves taking averages.

There's no modeling involved. You know, you don't need any fancy statistics to guarantee that the experiment gives you an estimate of the average causal effect of treatment.

Now, everything I've said probably seems so—some of you are real familiar with this having courses. Those of you who haven't—probably both groups are saying this is so simple; this is almost stupidly simple. I can't believe this guy is spending time, you know, telling us all this stuff. We already knew it.

But it's deceptive. One of the things that's deceptive about it is, I already embedded an assumption in what I told you that seems natural, but it can be wrong. Why are there only two possible outcomes? What if the treatment that I get affects your outcome under treatment? Or your outcome under not treatment?

The assumption that I slipped in here without mentioning 'til now is the "no interference between units" assumption, and people in experimental design have written about this. I mean, the Cox book is a book on experimental design that really predates some of the—at least the Rubin exposition of this kind of, this kind of causal modeling.

Sometimes, this is called the stable unit treatment value assumption, and this is another kind of unfortunate term, but the “no interference between units” label actually is more intuitive.

The point of my mentioning this is that the stable unit treatment value assumption—or the assumption of “no interference between units” —can be wrong. And an obvious case in which that can be wrong is that, is the example of vaccines and the response to vaccines. Think about this.

The response to a smallpox vaccine or not depends on who else is vaccinated. The reason we can eliminate—eradicate—smallpox from the world is because, at some point, the effect of being not vaccinated is not the same as the effect of being vaccinated.

And, as a matter of fact, when everybody is vaccinated is exactly the point at which the effect of not being vaccinated changes because you’re not going to die of smallpox without a vaccination.

The fact that, you know, the “no interference between units” can be wrong has been known for a long time among people who study vaccines and contagion of various kinds.

It’s also familiar to us. Consider the situation of classrooms or schools where we can have social interactions happening between the units. We worry about contamination if we were to assign people to different treatments within the same school, particularly to assign teachers to different treatments within the same school, and that kind of contamination is a violation of this stable unit treatment value idea.

Now, if it’s true that the effect of a treatment on an individual classroom depends on whether or not another classroom in that same school was assigned to treatment or control, well, that’s interference between units, and it’s one of the reasons why IES and researchers outside of IES, too, tend to favor a group assignment for some kinds of treatments where that kind of interference is possible.

Another point about causal inference and causal thinking, in general, is that some associations—some people would argue some associations—can't be causal, or it's not sensible to think of them as causal, and the primary case in which that occurs is when one of the two potential outcomes really is difficult to imagine existing because, remember, the whole premise of this sort of causal analysis is that there are these two potential outcomes.

If you have individuals who would never accept treatment no matter how they're assigned, if you have individuals who would always get in the treatment regardless of how they're assigned, if you have individuals who are defiers in the sense that they would always do the opposite of what you told them—you know, adolescents come to mind—then, it may not make sense to talk about causal effects on those units.

If you can’t imagine a unit getting treated or a unit not getting treated, that’s sort of equivalent to saying that potential outcomes don’t exist.

That’s what has led to this concept of compliers and the idea of the complier average treatment effect rather than the average treatment effect overall, to sort of exclude the cases in which you can’t talk about, sensibly about, causal inference.

On a more philosophical level, not all “what if” questions have causal answers—to borrow a phrase from Don Rubin. The idea of a randomized experiment—even if you’re not going to do an experiment—the idea of an experiment can help clarify what effects might be causal and what you really mean by causal effects.

And Don would argue, I think, if you can’t imagine an experiment that assigns people at random to treatments you’re interested in comparing, it’s probably not sensible to talk about causal effects of the treatment.

For example, it might not be sensible to talk about sex differences as a biological phenomena having causal effects because you can’t imagine the experiment of randomly assigning somebody’s sex as a biological feature.

It might be sensible to talk about gender as a sort of social reaction to biological characteristics as having causal effects because you might imagine an experiment in which people were gender-blinded or the apparent sex of an individual was altered. But, anyway, that’s in some ways a philosophical point.

But one thing that is important is that it’s not so clear you can talk about causal effects on the people who are never takers of the treatment, who are always takers of the treatment, or who are defiers in the sense that they always do the opposite of what you ask them to do. That logic has led to people limiting the scope of application of causal inference.

Now, what we all know is that randomized experiments are wonderful in the sense that they give model-free estimates of average causal effects. We know that randomized experiments are expensive, and they aren’t the right answer for addressing every question, and they particularly aren’t the right answer for addressing questions that are just at a more hypothetical stage, where we’re trying to get evidence that there’s a potential causal effect, and it’s worth the time and energy and money to carry out an experiment to investigate them.

That naturally leads to questions about, well, is there any other way to get either solid causal effects or plausible causal effects? Point number one is that there are really no other model-free methods known other than randomized experiments to get estimates of causal effects. Okay. What, we’ll take that as a premise.

There are many other methods that can give estimates of causal effects given that a model is true, and the key problem with all of these methods is that the model has to be assumed to be true—something that’s often difficult to verify empirically, sometimes impossible to verify.

But the whole point of these methods is to suggest plausible causal hypotheses that we can investigate, and I want to talk about three strategies that are very close to experiments but not exactly experiments.

Well, I guess the first—and to do that, the first thing that I want to do is just introduce a little notation here. And, whoops, pushed the wrong button. And here, suppose Y is an outcome variable, Z is an assignment variable, and suppose it's a dummy variable—you know, 1 for treatment, 0 for control, the natural way to sort of analyze experimental data or even nonexperimental data to estimate treatment effects might be to regress Y, the outcome, on Z, the treatment dummy variable.

If we do that, the estimate of the coefficient for the dummy variable turns out to be just literally the mean difference between treatment and control plus a difference in the average errors. If we sort of use this regression equation and substitute 1 in for Z, we see that the average in the treatment group is Beta-nought plus Beta 1, plus the average of the residuals in the treatment group.

If we substitute 0 into the equation, we see that the average score in the control group is just Beta-0 plus the average of the residuals in the control group, and so, the standard estimate of the treatment effect would be the mean difference, and, in terms of the model parameters, the mean difference would be Beta-1, which is the true effect of treatment, plus the difference in the average residuals in the two groups.

Now, when treatment is randomly assigned, then Z here is uncorrelated with the errors, and it turns out this is closely related to something about the potential values—potential outcome values—but we'll go into that later.

If the treatment assignment is uncorrelated with the residuals, then the average values of the two residuals, at least, you know, in large samples, are going to be the same, and, therefore, the treatment effect estimate, in fact, estimates the true value of the treatment effect.

This implies that with standard methods, OLS gives you an unbiased estimate of our Beta-1, our effective treatment, which we've already mentioned is the causal effective treatment.

Now, what goes wrong when we don't have randomization is that there is no guarantee that Z, the assignment variable, is uncorrelated with the residuals in the model, and, therefore, the treatment effect, the usual estimate of the treatment effect, estimates Beta-1, the treatment effect, plus this difference in average residuals.

And if the assignment is correlated with the residuals, if it's not exogenous, then we have the fact that the average values of the residuals are unequal, and so, the quantity that the analysis estimates as the treatment effect isn't the treatment effect; it's the treatment effect plus something.

And that means that the standard analysis you might do, even if you fancy it up a little bit, you know, may still give biased estimates of the treatment effect.

One approach to dealing with this situation is the instrumental variables approach, and many of you have probably heard about this—and I've actually talked about it at this meeting in previous years—and, in the instrumental variables approach, you make use of some more information. You make use of a so-called "instrumental variable," which I'm going to characterize as X here, and the basic idea of these two models together is that we want to know the effect of Z, which is assignment to treatment, on an outcome Y, but we don't have random assignment, so we make use of another variable X that has some useful properties.

The useful properties are that the instrument, the variable X, is correlated with assignment, so you can actually use the instrument to predict assignment, and, actually, I'll stop calling Z "assignment" and start calling Z "being treated" because this is the modal case in which you apply this methodology.

You have a variable Z, which is whether you actually got the treatment frequently, and you have a variable X, which is whether you were assigned to treatment. You know, this is the problem of estimating treatment on the treated effects.

If Z is actually getting the treatment, and X is being assigned to treatment, it’s pretty clear that X being assigned to treatment probably predicts whether you actually got treatment. Unless your experiment is really in bad shape, that ought to be true.

It’s also probably true that assignment to treatment is uncorrelated with the residuals in the top model. Another way of saying that is the only way assignment to treatment can have an effect on outcome is through whether you actually get treated or not, and, if that’s true, then some of the conditions for the instrumental variables analysis are satisfied, and I’m not going to say how the analysis actually goes, but it roughly, you know, exploits solving these two equations in a sensible way.

To be precise, one way to get the instrumental variables estimate is actually literally to regress the assignment—the did you get treatment variables, the Zs, on the assignment variables, the Xs, and then substitute in the top equation, not the actual assignment, but the predicted assignment, not the actual treatment received, but the predicted treatment received based on the assignment, and so, there's a standard analysis that goes with this model.

The reason I mention it is that among the class of quasi-experimental methods, this is one that can actually give you unbiased estimates of causal effects.

The famous paper by Angrist, Imbens, and Rubin showed that the instrumental variables analysis can estimate the average causal effects of getting the treatment on an outcome, provided this "no interference between units" assumption holds, provided that the instrument is randomly assigned, provided that this exclusion restriction holds basically that assignment affects outcome only through actually getting the treatment or not, provided assignment is actually related to getting the treatment, and provided that there are no defiers.

That’s important because it provides a different way of estimating causal effects than experiments. It’s not model free. These assumptions are actually tough to meet in a lot of cases, particularly the exclusion restriction. The argument that I’ve got something that predicts actually getting the treatment that can only have an effect on outcome through getting the treatment or not, that is an arguable and very difficult to verify assumption.

But the good news is that this is an example of a model-based method that can give valid estimates of average causal effects in the absence of randomization.

There is another strategy that is close to an experiment but not quite an experiment in the usual sense, and this is the idea of assignment to treatment with a probability that depends on the value of a covariate.

One example of how you can imagine such an experiment working is that if you have a series of kind of mini-experiments, where people are grouped according to some covariate value, you may assign people, randomly assign people to treatment or control, with the probability that depends only on the value of the covariate.

This can actually lead to designs that are actually pretty interesting designs. They allow you to, for example, if X is a sort of pretest covariate, something that indicates need for an intervention, you can assign more people to get the intervention at the highest levels of need and smaller fraction of people to get the intervention at lower levels of need, and designs of that kind are sometimes useful.

One of the things that is true about a design of this kind is that you can show with the same kind of logic that we showed before about the randomized experiment that conditional on the covariate, conditional on the assignment variable—what you’ve got is a little minirandomized experiment.

The way to think about all of this is you can either look at the conditional expectations, or I’ll just say what the logic is. The logic is for a fixed value of the covariate, each one of these, you know, you have a kind of mini-experiment. Conditional on X, you make a random assignment, not necessarily with 50 percent probability, to treatment or control so that conditional on X, for each value of the covariate—think of it as discrete, if you like—for each of the discrete values of the covariate, we have a little mini-experiment.

We can, in fact, estimate the average causal effect of treatment by estimating the local causal effect of treatment, so you estimate the treatment effect—the average treatment effect—for each value of the covariate, and then to estimate the average treatment effect for the entire population that is across the values of X, you just create a weighted average of the treatment effects, and it can be shown that that’s an unbiased estimate of the causal effect in the population.

There’s an important thing about this design that you should know, and that is that you won’t get the right answer if you just dump all the treated people into one group and the control people in another group and take the difference between the averages. That’s not an unbiased estimate of the average causal effect.

You have to look at the causal effect at each value of the covariate in this design and then add those things up, weighting in a way that’s sensitive to the representation of the values of X in the population.

Now, that design is very close to the regression discontinuity design, the RDD, and the regression discontinuity design, you can think of as being just like assignment based on a covariate, except it violates the principle that there are some people getting treated and some people getting control at every value of the covariate.

In the regression discontinuity design, the probability of treatment is essentially 1 above some cutoff on the covariate and 0 below some cutoff on the covariate. So, it violates the principle that what you've got is a lot of little mini-experiments. There is no real experiment going on anyplace here.

But, in the regression discontinuity design, you can show that you can estimate the local casual effect of treatment—local average causal effect of treatment—at the cut point.

The reason is, at least conceptually, that the RDD is almost a randomized experiment at the cut point. All the people just above the cut point versus just below the cut point are essentially, you know, that division is essentially random at an infinitesimal level.

Another way of saying this is that as the covariate value tends to the cut point, you've got a randomized experiment. So the sort of technical way of putting it is that in causal terms, the limit of the outcome for treated people as X goes to the cut point from the top minus the limit for people; the limit of the control outcomes for people who get control as they go from the—as they go towards the cut point from just the below side—is actually the causal effect at the cut point.

It’s possible with this design to estimate the causal effect at the cut point, but not every analysis can estimate that well.

The problem is that analyses that try to estimate the discontinuity—and the reason they call it a regression discontinuity design is that essentially the treatment effect at the cut point is how much different, you know, the control line is from—control regression line is from the treatment regression line.

The problem in actually estimating that gap is that we usually—the natural way to do it for people like me anyway is to use some kind of a model, and, as soon as you invoke the model, then the estimates of treatment effect are model dependent.

In principle, nonparametric regression

methods can provide model-free estimates of the causal effect of treatments in RDD designs, but the problem with these methods is they make their own set of technical assumptions about things like bandwidth and various smoothing parameters.

I think the best thing we can say is that the design is capable of yielding unbiased estimates of causal effects at a particular point, but the analyses that we use to actually get the estimates usually involve some kind of model dependency.

One other thing is that I've emphasized RDDs as having one cut point in estimating causal effect at a certain value of X, but you can obviously create regression discontinuity designs that have many discontinuities at various points and estimate causal effects at many different places on the X dimension.

Okay. I'm getting to differences in differences. You may not believe that.

[Laughter.]

DR. HEDGES: The nonequivalent control

group designs are really the workhorse quasi-experimental designs. And you understand how they work. They compare a treatment group with a comparison group that wasn’t randomized. There’s a huge range in quality of these designs, and even the people who advocate quasi-experimental designs would tell you that the range of quality is from pretty good to really awful, and that probably most nonequivalent control group designs in the literature are really awful, although there are occasional good ones.

One of the things that is true about these designs is they almost always rely on matching or adjustment for covariates, which is, kind of, you know, a statistical adjustment, is a kind of pseudo-matching. The question that many people are very interested in right now—my colleague Tom Cook and Will Shadish and a whole set of people are very interested in the question of whether quasi-experimental designs with a nonequivalent control group design, in particular, can they ever yield causal estimates that are close to the right answer? It seems clear that they can, but the estimates are never model free.

The problem with nonequivalent control group designs is essentially the fundamental problem that you get with nonrandomization. The question is, “Do you have what a nonequivalent control group design is trying to do if it’s trying to make the data ‘as if random’ assignment controlling for covariates or subject to matching?”

If the design succeeds in making the treatment assignment “as if random” controlling for various other things, then it will give you the right answer. If it doesn’t, if the analysis doesn’t succeed in doing that, it will give you the wrong answer, and it’s never very easy to tell whether or not you’ve succeeded in creating the conditions necessary for making the treatment assignment essentially “strongly ignorable.” If we think of this in the missing data language, “as if random” is the same thing as “ignorable.”

If we get to the stage where we’ve controlled for enough, the treatment assignment is essentially random, of course, the analysis will give us the right answer. The problem is you never know when you've quite gotten there.

I think I'm going to—well, I'll say a word about this. It's easy to write down conditions in which the nonequivalent control group design will give you the right answer of the average causal effect. It all depends, of course, on what the response functions are. In other words, what these potential outcomes look like as a function of covariates and other things.

It's easy to write down models in which the design can give you the right answer. If, for example, the relationship between the response and covariate X is essentially linear, as I've written here, and the relation between the covariate X and the response among treated people essentially follows the same regression except that there's a constant shift, then it's easy to show that, you know, most analyses will give you the right answer.

You'll get an estimate of Tau that's unbiased pretty easily.

But that’s only one form the response functions can take, and we don’t necessarily know what the functional form is. If you change the response function a little bit, if we say—and the only thing I’m going to change is have a different slope for the response function for untreated and the response function for treated—if the covariate dependency is just a little bit different in those two, then it follows that the usual estimate of the treatment effect, the OLS estimate, will be the actual causal effect of the treatment plus some average of the slopes times the difference in the covariate values between treated and control.

It doesn’t take much very tweaking to a model that gives the right answer to produce a model in which the analysis will give the wrong answer, and that’s the problem with model-dependent estimates.

Now, what you may be saying is, “Well, I see how to fix that up. We could do an analysis that would fix up the situation and remove the bias.” Well, but that’s exactly the point. To get an unbiased estimate of the causal effect, you have to know what the right model is, and, therefore, your analyses depend on the prescience of model specification.

It’s not easy to know what those models might be—what the right models might be. And I picked an example that was really simple, and I could have made the example a lot more complex, and it would be obvious how difficult it would be to distinguish the right answer from the wrong answer.

Okay. Now, I’m getting to the part that was promised, and that is differences in differences. The difference-in-differences idea can be seen as a particular kind of nonequivalent control group. I mean, it’s, yes, economists use it, but that doesn’t mean it doesn’t fit into some framework for thinking about designs that you all know. It fits into a slot in the quasi-experimental design world that’s well known to you.

Now, difference-in-differences is used often to evaluate effects of policies in education or elsewhere, and the way to think about differences in differences is that assume that there’s a series of longitudinal observations in various locations, like states where a policy has been implemented at some time and in some locations, maybe not in all locations, maybe not across an entire state, but maybe across part of a state. We can identify which people are subject to the policy and which aren’t.

And crudely, what difference-in-differences analyses do is they estimate the effect of a policy by comparing the difference in outcome before and after the policy is implemented on the individuals to which the policy applies.

Then we look at the same difference for individuals to which the policy doesn’t apply, and the logic here is, well, if the policy makes a difference, then we’ll see a change between before policy and after policy for the people to whom the policy applies. We won’t see a change for individuals for whom the policy doesn’t apply or at least whatever change we see for them is not a consequence of treatment.

And so by comparing these differences, we’ll get an estimate of what the policy effect—the treatment effect—is, and that’s why they call it a difference-in-differences estimator. It’s essentially taking the difference between, if you like, posttest minus pretest among those who got treated and posttest minus pretest among those who didn’t get treated.

That’s the absolute simplest version of this. Usually, the difference-in-differences estimators are set in a more complex and elaborate analytic scheme that is actually more convincing—certainly more convincing to economists and more convincing to statisticians.

To give you an example of a sort of very typical setup for difference-in-differences analyses, suppose Y_{ist} is the outcome for individual i in location s—you might want to think “states” for locations—at time t.

What we’re going to have is a whole bunch of Y_{ist} values for many individuals in many locations across a long time span. Then, we can have an X_{ist} that’s a corresponding individual level covariate used as a control variable in the analysis, and then the analytic model that might be invoked is one in which we say Y_{ist} looks like a fixed effect of locations plus a fixed effect of time plus effects of covariates plus a sort of a treatment effect, and Beta here is going to be the treatment effect we tried to estimate, and T is a dummy variable for whether an individual in a given location gets—at a particular time, is subject to the policy or not. And then Epsilon is the residual.

This is one analytic scheme. Usually, there’s a sense that there are important differences between locations and important differences between times that need to be taken into account, and you can either do that by what the economists would call doing a fixed-effects analyst—making a fixed-effects transformation, putting dummy variables in for those things.

You could also do it via random effect strategy, like centering or just modeling as random effects. But forget about the analytic details for now. The fixed-effect strategy is understandable, and it’s perfectly suitable for talking about this, and it’s actually probably what most of these difference-in-differences analyses do.

Now, one thing that you would all—many of you would recognize about a model like this one, if we’re talking about individuals who are located in various states, say, this error term here is, they’re not—individuals are not—error terms are not all going to be independent of one another. There’s a clustering problem here that you got to take into account. If you don’t take the clustering into account, you have big troubles with this analysis.

If you think that people in the same state or the same location are more similar to each other than people in different locations, that induces clustering. So that has to be taken into account. That’s a technical detail.

The idea of difference-in-differences as an estimation strategy has great appeal. If you have good longitudinal data sets, it’s easy to use, and that’s why a lot of economic analyses have used it based on good economic longitudinal data sets.

But difference-in-differences has another virtue. It’s easy to explain to policymakers. It’s easy to understand yourself, and, as important, if you’re trying to advocate policy based on some statistical analysis, you got to be able to explain it to policymakers. Experimenters, you know, like experiments because we can explain them to people, and people understand them.

We don’t have to say, you know, there’s this complicated model, and if you take all these things into account, blah-blah-blah. You can say something that’s almost as simple as the experiment with differences in differences, and that’s a great appeal.

In some ways, it seems to be the natural analysis to learn from natural experiments, where a policy has been tried in some places and not in others or tried at some times and not in others. So, there are plenty of reasons why people are interested in differences in differences.

It may seem that the model I just gave—with all the fixed effects and everything—is going to be hard to formulate in causal terms, and I agree, it’s a little tricky to do, but the difference—but the way to think about it is—that the difference-in-differences analysis identifies the treatment effect by looking at the pretest/posttest change for those who got treated versus the pretest/posttest change for those who didn’t, and that boils down in a technical sense to saying—well, that leads us to saying that what the difference-in-difference analysis estimates is this difference for the treated group minus the difference for the control group and—conditional on treatment or control.

The thing is, this is not the causal effect of the treatment. If you make the right set of assumptions, this becomes a causal effect of treatment, but I write this down basically to say to you when you think about it carefully, it’s pretty clear that difference-in-differences does not automatically estimate the causal effect of the treatment.

Other things have to be true. Other assumptions have to be true for that to be the case. Again, if you—it’s easy to write down a response function. It’s easy to write down a relation between treatment and response so that difference-in-differences will estimate the right answer.

But the question is, does, you know, are the conditions that are needed, namely, that the difference between pretest and observed outcome for the treated is an estimate of the difference between pretest and observed outcome for—would be the same as if it would be the same individuals had been assigned to control and vice versa? It’s the same idea of whether assignment is independent of—is independent of the two key differences that go in the difference-in-differences analysis or not.

You can think of many cases where this is not going to be true. If the actual pretest causes both the policy and is correlated with the outcome, then the necessary condition isn’t going to be true. It can be something else that causes both the preoutcome, the preintervention outcomes, and the assignment—the decision to make a policy.

In other words, what I'm saying is if the thing you're observing—if the outcome you're observing over time—actually causes the policy to happen, then the difference-in-differences analysis is not going to give you the right answer, and this is sort of the fundamental kind of endogeneity problem that we encounter in a lot of different settings.

Now, there are informal checks about this. You can look at trends beyond the time of policy implementation and see if you can convince yourself. You can estimate the effects of treatments where there has been no policy change as a check, and this is very convincing. It's a very convincing deflation of a difference-in-differences estimate if you can show that if you had picked three years before the policy was implemented and done the same analysis, you get the same treatment effect. That would not be good.

[Laughter.]

DR. HEDGES: This is an informal way that people go about checking, but these kinds of checks are not—they're suggestive—they're not definitive. They can invalidate the analysis, but they can't convince you the analysis is right.

Anybody who's interested in doing difference-in-differences analyses should plan on devoting a great deal of their time to carrying out these kinds of checks that could invalidate the analysis as a way of just protecting yourself against being really embarrassed when you get up to talk about your results in front of a room and somebody may have done that check and discovered that your analysis doesn't pass muster.

I think beyond the kind of analytic checks of looking at trends in the data, looking at what would happen if I had picked a different point in the time series to evaluate the treatment effect, there's also sort of conceptual analysis. What do you know about this? What do you know about the policy environment in the places where the policy has been implemented? And that can help, too.

But there’s a smaller problem associated with difference-in-differences analyses that you also ought to know about, and when I say it’s a smaller problem, I don’t mean that it’s not enough to completely invalidate the analysis. I just mean that it’s one that has more—it’s more amenable to technical solution.

That is, that remember the kind of data you tend to use in difference-in-differences analyses are long time series. You have the observed outcome over a long sequence of time points, and then there is some point—I’m speaking a little bit metaphorically here, but not entirely. You look at the outcome over a long series of time points, and then there is some place where the policy is implemented, and you say, “See, that change is a lot bigger than the changes everywhere else, and it’s bigger than the change in these folks in this place where the policy wasn’t implemented.”

There’s a kind of a logical desire to have a long time series so that you can show how outstanding the change is at the place where the policy is implemented.

It turns out that that—that sets you up for an interesting analytic shortcoming. There are two things that are true about analyses based on time series like this. If they have—first of all, the kinds of outcomes we tend to trace over time—policy outcomes, we tend to trace over time—often have very high autocorrelations. There's very high correlation between one time and the next.

The second thing is that the policy variable, the independent variable, the dummy variable of policy or not, also tends to be very highly autocorrelated. Think about this for a minute. Suppose we have a 16-year time sequence, and we are looking at test scores every year, and then, at one point in that sequence, there's a policy change, and we want to evaluate the effect of the policy change.

Think about the independent variable of the dummy variable for treatment. The dummy variable for treatment is zero, zero, zero, zero, zero, zero, zero, until you get to the point where the policy has been implemented. Then it’s one, one, one, one, one, after that.

If you think of the independent variable, it has very high autocorrelation. The correlation between the value at one time and the value at the next time is perfect, all except for the place where there’s a jump.

It turns out that if you ask the question, “How does, how do autocorrelations affect the results of OLS analyses, standard analyses,” the answer turns out to be that positive autocorrelations tend to mean that the standard errors—the uncertainty of the treatment effect estimates—are too small. In other words, you underestimate the uncertainty of the treatment effect estimate.

It also turns out that everybody knows that, but it’s sort of less well known that autocorrelations among the independent variables have an impact on the standard error of estimated treatment effects, and the higher the autocorrelation there, the higher the overestimation in precision of the treatment effect.

And the longer the time series, the more both of these things impact the estimate of the standard error. And, you know, this is a mathematical fact, but it's also something that makes perfect sense.

What does it mean when you have observations that are correlated? Well, it means there's less information there than there is if they weren't correlated. So, these autocorrelations lead to dramatic underestimation of the real uncertainty of the estimated treatment effects.

When I say that this is dramatic effect, I don't mean, you know, you get—that the real significance level is .06 when it should be .05. I mean, the real significance level can be .4 when you think it's .05, or .6 when you think it's .05.

This, although it's a small problem, it's a small problem that can pretty much invalidate the analysis just as much as ignoring clustering can invalidate the analysis.

There's an awful lot of published work that hasn't taken, you know, that hasn't addressed this problem at all—published work by good people in good journals, as a matter of fact.

The standard error problem I've just been talking about is difficult but not really impossible to solve. One way to go about it is by using generalized least squares analyses. This can be done, but inference for the autocorrelation is usually not very good. This is not a perfect answer, but it's an answer that is a heck of a lot better than what people usually do.

Another answer is that you can use a variant of robust standard errors, which works pretty well provided you've got enough locations to have a large sample of locations. We don't always have that in difference-in-differences analyses.

Randomization tests seem to work well in problems like this, and, by the way, this—all of this stuff—has a lot in common with the problems of single-subject designs. It’s really the time series aspect of both problems that respond to the same set of statistical solutions.

Sometimes, collapsing the data into just two time points—you know, before and after treatment—and analyzing them can improve performance of analyses.

There is not a completely simple answer to how to deal with the standard error problem that solves all of the issues. Probably the thing that comes closest is the use of robust standard errors, if you have a lot of locations, and, if you don’t, then trying to parameter—then trying some of these other strategies may work.

The thing I want to leave you with is that difference-in-differences is an interesting strategy. It’s an easy-to-use strategy if there’s good data, and it can be suggestive, but, without randomization, causal inference is really much harder and much more model dependent than it is with randomization.

And so, even a technique as appealing as difference-in-differences has to be scrutinized very carefully. You have to do a lot of sensitivity analyses, and, even then, you need to be very careful about what you conclude. But, if you think of it as a technique for generating suggestive causal hypotheses, there is probably some real virtue in it because the analyses are relatively easy to do.

Now, having said that the analyses are relatively easy to do, you probably have to do a few hundred analyses to convince yourself that it isn’t so sensitive as to be unbelievable. So, I’m not sure that I ought not to qualify the statement, “It’s really easy to do,” by saying, “It’s really easy to do one difference-in-differences analysis.”

To do all the sensitivity analysis that probably should accompany it, like, for example, choosing time points at random and seeing whether or not you get intervention effects there, that is quite a bit more complicated. And, if you have to do randomization tests, that’s even more complicated.

Having said that, I think I will now stop, and I won’t stop just yet. I will say there’s—I put some references on the technique, particularly differences-in-differences technique, in my slides here, not because I expect you to write them down, but I suspect these slides will be made available someplace.

In particular, the Bertrand, Duflo, and Mullainathan paper is—actually, I commend it to you. It’s very readable, and it’s also quite good. I’ve got a set of references here, and, beyond that, I just will thank you all for enduring this unnaturally long talk, and I’ll let Allen do whatever he wants to do now.

[Applause.]

DR. HEDGES: Do you want me to leave?

DR. RUBY: No.

DR. HEDGES: Okay. If there are questions, I’ll take them. If you all just want to go home, that’s okay, too. Well, it seems like people want to go home—oh, no, wait, wait. There’s a lamb to the slaughter here, good.

Go ahead. Maybe go to the microphone.

MR. VAN HOUDNOS: My name is Nathan Van Houdnos. I just finished my first year in the statistics Ph.D. program at Carnegie Mellon, so I'm going to ask you what might even be a dumb question.

DR. HEDGES: They're usually not, actually, in my experience.

MR. VAN HOUDNOS: When you looked at all these different approaches, you made a big point of making sure that they were unbiased estimators.

DR. HEDGES: Yeah.

MR. VAN HOUDNOS: I was wondering if there are other methods that will allow for some bias if it squashes the variance of your estimators so that you can sort of approach it in a more mean-squared error kind of way?

DR. HEDGES: The answer is surely "yes," and I can probably think of some, but I think the bias issue is a big deal because the bias usually dominate—in—well, the thing we're really worried about—I appreciate this problem because there are other settings in which, you know, I’ve argued just what you’re arguing—that don’t worry about unbiased because the variance is so big, it dominates.

But, in a lot of these cases, I think the problem is that it’s the bias term that dominates the mean-squared error term. That the precision is already small, and, in fact, that makes the bias—and, in effect, that makes the bias a bigger problem.

When you have a lot of variance, you can tolerate some bias because as long as it’s not big compared to the variance, it doesn’t mislead you. I think the problem with a lot of social experiments and other analyses—like analyses of big data sets with people from all 50 states, to give an example of how difference-in-differences is often used—the estimated precision is really high, and a tiny bias just swamps all of that, all of that variation.

It’s still true that we want to get the variance right because frequently we’re comparing a small effect with a small variance, and the example in difference-in-differences is if you get the standard error wrong by 30 percent, even though it’s an absolutely small number, it may make a huge difference in the rejection rate.

The answer is “yes,” it’s important to consider variance, but, in a lot of cases, it’s the bias that really is the driving term of the mean-squared error.

MR. VAN HOUDNOS: Great. Thank you.

DR. HEDGES: Yeah.

DR. SCHMIDT: Hi. I’m Stefanie Schmidt. I’m with IES. I was wondering if there’s a particular robust standard error estimation technique or a set of techniques that you would recommend for difference-in-differences?

DR. HEDGES: It boils down to—yeah. It boils down to essentially taking the location. Of course, now, I have to say it all depends on the details of the model, but, in the example I gave, which is a fairly typical setup where you have—let me see if I can go back to that—in this setup, where you have fixed effects for—oh, that’s another—where you have fixed effects for locations and times and then a bunch of individual level covariates, the natural robust standard error approach would be to treat the locations as clusters so that in the case of an analysis of 50 states, then you'd have states as clusters.

Therefore, a question that a reasonable person could ask is, "Is 50 states enough for the robust methods to work well?" And it's probably on the borderline. You know, it's probably—and a lot of people say, "Oh, it's just over the borderline, you know. Fifty is enough."

But, if you had analysis with 20 schools, then I'm not sure that I would say 20 schools is enough, or 20 school districts, so that's the bind you get into.

But the basic idea is if you were to use a standard kind of approach and treats locations—my S's—as clusters, that's the kind of approach that seems to get pretty good, pretty good Type I error rates.

I should add—I could add that it isn't actually just that you get the significance wrong.

The estimates seem to be inflated as well, so it’s a little more than just you don’t get the rejection rates right with standard methods. You don’t get the estimates right either.

DR. BERNSTEIN: Hi, Larry. I’m Larry Bernstein from RTI International. When we’re analyzing experimental data, there’s often debate amongst people who think you should estimate gain scores or you should estimate mean differences and control for a pretest, or some people even say you should estimate gain scores and control for some other covariate or a pretest.

And I wonder whether this same discussion is also applicable to your discussion when you’re looking at, you know, nonequivalent designs, and if there are any apparent solutions to that question?

DR. HEDGES: Aah. Well, there I think if you’re in the realm of—I’m a little bit like a broken record on this—if you’re in the realm of randomized experiments and pretty good randomized experiments, then lots of these—the problems that arise in quasi-experiments, you’ve really gotten—

you’ve got clear of the biggest kind of problems.

But the problem you’re talking about strikes me as much more of a conceptual problem of what treatment effect you want to estimate. It’s—a famous paper I’m sure you’re familiar with and maybe a lot of people are, too—but I’ll mention it because it’s relevant here—is the paper that Fred Lord wrote which describes what’s come to be known as Lord’s paradox.

Lord’s paradox can be described pretty simply in the following way: that—and this is how Fred Lord described it—somebody running a university has, you know, has just—is interested in evaluating the food service, and they measure the weight of all the individuals who come into the college as freshmen, and they measure the weight at the end of their freshman year, and they want to know what this means about the effectiveness of the food service, and, you know, you got pre/post design, you know—not the strongest design in the world but at least might be good enough to learn something from.

And so, they give the data to two statisticians, and one of them, one of them computes the gain scores and discovers that the average, that on the average, neither the males nor the females have gained any weight over the year.

And another one, the other statistician, does another perfectly reputable analysis. They look at the effect—the gender effect on weight—controlling for pretest. And they get a different answer. Both answers are right, but they're answers to different questions.

It strikes me that something we haven't been really good at in education research—but I'm not sure people have been particularly good at it anywhere else either—is thinking about what question you want the answer to from your analysis.

If you're interested in the causal effect on gains, you know, Y minus X type gains, then that's what you ought to analyze. If you're interested in the causal effect on an outcome controlling for as if everybody started out at the same place, that's a different question—a subtly different question—and the analysis of covariance answer is the right answer—is the right way to get the answer to that question.

But I think that it's at the conceptual level where we really need to think, and it's really hard to think about because most of us aren't really used to thinking about even that there's a difference between those two questions.

I think that's where in the experimental world we really have to think a lot, you know, about which, question do we want the answer to, and then make sure we can relate the analysis to the answer we want to get.

I mean, it's sort of funny because it's an elementary blunder in a lot of the proposals that people write. You know, you say, oh, well, the analysis doesn't have anything to do with the question they're asking.

But there are subtle variances of that that I'm sure we all fall into—I'm sure I've fallen into it—and I don't know much about what—I don't know much about what to say except that it really is trickier than it seems and requires more thought and more smart people sitting around a room before you go off and do something—thinking about it—than most of us are willing to give it.

We're about one minute to go, so we could either quit or I'll take one more question. I'll quit.

[Laughter and applause.]

[Whereupon, at 11:47 a.m., the panel session was concluded.]