Institute of Education Sciences

Fourth Annual IES Research Conference
Concurrent Panel Session

"Problems with the Design and Implementation of Randomized Experiments"

Monday
June 8, 2009

Marriott Wardman Park Hotel
Thurgood Marshall South
2660 Woodley Road NW
Washington, DC 20008

Contents

Proceedings

DR. RUBY: Good afternoon, everyone. I am Allen Ruby from IES. Thank you for attending, and just a reminder, if you ask a question, because this session is being taped, please state your name and your affiliation if you'd like to.

Last year, I introduced Dr. Larry Hedges at the IES Research Conference. He was doing a talk on instrumental variables, and at that time, I noted he had joined the Northwestern faculty in 2005. He has appointments in statistics, education, and social policy. He is one of the eight Board of Trustees Professors at Northwestern.

Previously, he was the Stella Rowley Professor at the University of Chicago, and his research has involved many fields including sociology, psychology and educational policy.

He is widely known for his development of methods for meta-analysis. He has authored numerous articles and several books. He has been elected a member or a fellow of many boards, associations, and professional organizations including the National Academy of Education; the American Statistical Association—come on down. There are plenty of seats up front, please— American Statistical Association; the Society for Multivariate Experimental Psychology.

He has served in an editorial position on a number of journals including the *American Education Research Journal*; the *American Journal of Sociology*; the *Journal of Education and Behavioral Statistics*; and the *Review of Educational Research*.

So I don't want to go into the details again this year. I really just want to ask what's he done for education research lately?

[Laughter.]

DR. RUBY: And to that end, I'd like to note two areas of his current work that might be of interest to some of the folks here.

When we evaluate educational interventions in schools, we often act as if those schools included in our sample are a probability sample from a population of schools, when, in fact, they're really not. They're usually a convenient sample of schools that are willing to work with us.

And so Larry is currently working on how to address the nonprobability sampling in order to improve our estimates of population treatment effects or to improve the generalizability of the results from our experiments.

He has published one piece of this work in a chapter entitled "Generalizability of Treatment Effects," in the book Scale-Up in Principle, 2007, edited by Schneider and McDonald.

A second area is as we move to a greater use of cluster randomized trials, there are implications for what the effect sizes are and mean; and he has been working on improving methods for defining effect sizes in cluster randomized trials; and he has published two pieces of work in this area: one in 2007 in the *Journal of Educational and Behavioral Statistics*; and another one in the recently released *Handbook of Research Synthesis*; and another piece will be coming out soon in JEBS on three-level designs.

Third, and what he's talking about today, he's been considering a set of issues that often crop up when designing and carrying out randomized trials. These issues have important implications for preserving the integrity of the experiment, doing the analysis, and interpreting the results.

So I'd like everyone to please welcome Dr. Larry Hedges.

Thank you.

[Applause.]

MR. HEDGES: Well, it's always tough to live up to an introduction like that, and I'll just begin this talk by saying that IES put me up to it, and they put me up to it in the following way: that some folks on the IES staff, and that actually wasn't just Allen, said to me, you know, we keep getting these questions from people that are really sort of simple questions, and I'm surprised that folks don't know the answer to those questions. Maybe you could just give a talk to clarify, you know, these things for people.

And that's how I got put up to doing what I'm going to try to do today. And when I was asked to sort of gin up a title for the session, I came up with the title "Problems with the Design and Implementation of Randomized Experiments."

But a few days ago when I was talking to Allen about this session in somewhat greater detail, he helped me to realize that a better title for this talk would be "Hard Answers to Easy Questions."

[Laughter.]

MR. HEDGES: Because the questions that kept cropping up were in principle easy questions, but in order to give a good answer to them, it actually required thinking through quite a few things that weren’t necessarily obvious either to the person asking the question or to the person answering the question.

I think I know, I see a lot of my methodologically inclined friends in the audience, and I know we’ve all had the experience of somebody coming up to you and saying I’ve got a really simple question, and they think it has a yes/no answer, and they think that enough information to answer the question is contained in the first simple phrasing of it, and then we have to sit them down for a few hours and talk through a lot of details of what they actually mean, and so this, in a sense, this talk has some of that character.

I’m going to start with an easy question that has arisen frequently in my experience, and I guess in IES’s experience as well. People will say isn’t it okay if I just match something like schools on some variable before randomizing? And then they often go on to say, you know, lots of people do it.

[Laughter.]

MR. HEDGES: And this question has been asked of me directly by some fairly sophisticated people, and, you know, when they say that, you know, “lots of people do it,” what they are really saying is, “come on, it can’t make that much difference, and I know lots of people who do this and kind of ignore the fact that they did it later. You know, surely, this is okay; right?”

And I see this as an example of a simple question, but giving it an answer requires some serious thinking about design and analysis, and in a certain sense, this is an opportunity to raise people’s consciousness about some things you probably already know.

Maybe for many of you, all of the things are things you already know. In that case, you don't need this talk, but if there are some things that you might not have thought about quite this way, maybe it will be helpful.

Well, the first step for me in thinking about a question like this is trying to understand exactly what the question means, and not the only way, but one way to think about it is that adding a matching or a blocking variable means adding another factor to the research design.

So what I'm going to do today, I'm going to rely heavily on kind of a standard analysis of variance, standard experimental design approach to thinking about these things, not because analysis of variance is necessarily how you would analyze the data from a real experiment, but because it makes certain things really clear and has a sort of universal vocabulary that may be helpful.

So if you see a lot of analysis of variance kind of approach to things, that's why it's there.

And then to move on, that simple question, you know, has consequences that depend a little bit on which design you started with. You might have started with an individually randomized design, the sort of completely randomized design, no clustering or anything like that.

I want to talk a little bit about that because it's kind of the paradigm for answering the questions with respect to more complex designs that we may actually be more likely to use in education; things like cluster randomized designs, which might also be called hierarchical designs if you read a book like Kirk's *Experimental Design* book; and multicenter or matched designs, which might be called generalized randomized blocks designs if you look in a standard kind of experimental design book.

So if we started with, if the place we started was with an individually randomized design where we basically assigned treatments to individuals at random; adding a blocking factor; using the blocks; matching a little bit; corresponds to essentially changing the design from a simple individually randomized design to a randomized blocks design or what is actually usually called a generalized randomized blocks design.

And you could, and this is a sort of sample depiction, and I'm going to make our discussion easy by imagining that we've picked the levels of our blocking factor in such a way that we get sort of equal numbers and miraculously equal numbers in both treatment and control in each block.

So the first point that's useful in thinking about what happens if we had a blocking factor is that it depends on the design and that adding that changes the design.

Now, if we want to ask what the impact on the analysis might be, well, we could, again, think about this in analysis of variance terms, and if we imagine that we have gotten incredibly lucky in the choice of our blocking, levels of our blocking factor, so that we get equal numbers of students in each block, 2n students in each block, n in treatment and n in control, and p blocks, then we would have a standard partitioning of the sums of squares, and that goes, that we talk about in analysis of variance.

And I think of the partitioning of the sums of squares in terms of partitioning of the variability of the data, and so the notion is the total amount of variability gets partitioned into part that's due to treatment and part that's due to within treatments in the original partitioning.

And the degrees of freedom get partitioned, and the original test statistic compares the sum of squares due to treatment, which is a function of mean difference, to the variance within treatment groups, in other words, the sum of squares within treatment groups divided by the degrees of freedom there.

Now, by introducing this blocking factor, we've introduced the possibility of partitioning the variation further. If we start with our individually randomized design, but now introduce a blocking factor, we have a different partitioning of the total variation, one in which there is a certain amount of variation between blocks, a certain amount of variation that is associated with the block treatment interaction, the fact that within each block, we have both treatment and control individuals. So there's a block-specific treatment effect that we could compute.

And the block treatment interaction corresponds to an amount of heterogeneity that has to do with heterogeneity in treatment effects.

And then there's a sort of within cells, that is within blocks by treatments component.

Now, once we realize we have this new design with a new partitioning of the variation, the question arises: well, what's the right test statistic?

And I sort of mean this in two senses. You know one sense you could mean it is how do I actually compute the test, but also, you know, how is it sensible? With what is it sensible to compare the variation due to treatments in order to decide whether the variation due to treatments rises above the level of the background noise?

Well, that depends on the inference model. Now, that term may be—I'll just—this is just a review, by the way, of the sums of squares, that in essence, what used to be called within treatments in the original design now gets broken up into three components: a between blocks component; a block treatment interaction component; and a within blocks by treatments, a within cell component.

And that's an important thing about this design, a potential advantage of this design. But as I was about to say, in order to figure out what the right thing to do, now that we've added this blocking factor to our design, we have to think about the inference model that we want to attach to this design. And I need to say a little bit about that because one of the things that isn't discussed in quite these terms in many experimental design courses is the notion of inference models.

It's always talked about, but not exactly in this way. And I'd like to make a distinction between two kinds of inference models, one which I'm going to call conditional inference model, and one which I'm going to call the unconditional inference model.

And the point of making this distinction is that the inference model is associated with, determines, and is determined by, the type of inference that you want to make from the experiment; and once you've chosen the inference model, that has implications for the appropriate statistical analysis procedure; and to put it, to relate it to something else that probably confuses a lot of folks in their first experimental design course, the inference model determines what are the natural random and fixed effects in the model.

So let me say a little bit more about what I mean by inference models. The conditional inference model is one in which the generalizations you hope to make are to the blocks that are actually in the experiment or ones that are just like them.

And what does “just like them” mean? It means they respond, that in a sense they have exactly the same effect parameters.

In the conditional inference model, blocks in the experiment are the universe you want to generalize to. And the idea of generalization to other blocks depends on extra-statistical considerations. You have to make an argument that the results of this experiment apply to some other blocks because those blocks, those schools—frequently blocks are schools or they’re school districts—are just like the ones that are in the experiment, and that’s an extra-statistical argument.

So it’s clear that generalization from a conditional, in the context of a conditional inference model to things outside the observed blocks, say the observed school districts, is something that can’t be done in a model free way.

By the way—well, let me go on and say a word about the unconditional inference model. So what’s the alternative? Well, the alternative that I’m posing, and it may not be exactly the only alternative, is that we have an unconditional inference model, which means that the inference is intended to be a generalization to a universe of blocks which explicitly includes at least some blocks that are not in the experiment.

Therefore, we think of the blocks in the experiment as being a sample of blocks from a universe or population of potential blocks. There's space either place. And the idea, the usual idea, in the unconditional inference model is that the blocks in the experiment are a representative sample of some population, and the inference that's of interest is to the population of blocks that the experiment are a representative sample of.

And the beauty of the unconditional inference strategy is that inference to blocks not contained in the experiment is by sampling theory, and we have this wonderful theory of generalization based on sampling.

Now, if the blocks aren't a probability sample of any population that you can specify, the generalization gets tricky because the question becomes what is the relevant universe and how would you know; and, in a sense, extra-statistical considerations, just like in the conditional inference case, come into play and are, in some ways, just as tricky.

Now, I want to refer this situation that I'm talking about to a situation that you're all familiar with but may not seem exactly the same to you, and that is the case of stratification and clustering in sample surveys.

If you think about the way in which we do inference in sample surveys, many sample surveys pick intact groups, a sample of intact groups, clusters, and we get the sample by sampling first the clusters and then individuals within the clusters.

Now, when we do this, what we're doing is taking a sample of clusters and trying to make an unconditional inference to the universe from which those clusters have been sampled. If you happen to have all of the clusters in the universe you want to infer to, we use a different name for the clusters in sample survey work; we call them strata.

And when you have a sample that's based on strata, that is all the subdivisions of the population, and not clusters, some of the subdivisions of the population, it has implications for the inference. It has implications for the uncertainty of the inference, in particular, and it's exactly parallel here in the experimental situation.

So you can think of the inference model as being linked in an inextricable way to the sampling model for blocks. If the blocks that you observe are ideally a random sample of blocks, they're a source of random variation. If the blocks observed are the entire universe of relevant blocks, they are not a source of random variation. Just like clusters are a source of added variation in a sample survey but strata aren't.

Now, the reason I'm making a big deal about inference models, and I haven't said anything about statistical analysis yet, is that I think it actually—these two things are often conflated in people's understanding of what to do in experiments, and I think it helps to separate these two conceptually different things.

There is this inference in sampling model, and then there's what analysis you do to try and learn something from data you collect under those models, and the fact is you can do any statistical analysis with any sampling model; some of them won't be sound. But, nonetheless, these two things are in principle separable.

So going back to our simple example, if we think of the blocks as fixed effects, if we're willing to, if we're willing to entertain a conditional inference model, that is we're interested in inferring to the universe of blocks we have observed in the experiment, then it's entirely appropriate to think of the blocks as being fixed effects, and in that case, we get one test statistic which basically is the one that I've listed here as F conditional.

And the only thing that winds up in the error term in the denominator we compare to the variation due to treatment for the purposes of seeing whether the treatment variation rises above the noise is just the within-cell variation.

And there's the exact distribution of this thing is known. It's an F with the degrees of freedom that I've given here.

But if the blocks are random effects, in other words, if we think of the blocks we have introduced as not being a universe unto themselves, but being a sample from a universe we want to infer to—for example, we have school districts. We've blocked by school districts. We're not interested in making a generalization to the school districts we happen to observe. We want to make a generalization to the universe of school districts from which we could putatively say ours are a representative sample, then the appropriate test statistic is different.

And in this case, the appropriate error term is determined by the block treatment interaction variance, the sum of the squares is due to the block treatment interaction, and it's notable that that error term has a lot fewer degrees of freedom.

By the way, one of the ways of thinking about why this F unconditional is the appropriate test statistic is to sort of go back to the design and remember that in this design, we can compute a treatment effect within each block, and what this analysis of variance test statistic is, is basically just the equivalent of doing a t-test on the block specific treatment effects, only it's an F test. You know, square of a t-test on the block specific treatment effects.

Now, you can see, if you compare these two, that the error term in the conditional inference test, the so-called "fixed effects model," has a whole lot more degrees of freedom than the one in the unconditional model.

If p is the number of blocks, we have, you know, basically 2pn minus 2p degrees of freedom under the conditional analysis and only p minus 1 under the unconditional analysis.

And that's a difference so you can already tell the conditional test is going to be more sensitive.

What you can't see immediately from just comparing the test statistics is that if there is a treatment effect, the average value of the F statistic is going to be larger under the fixed effects model. Generally, it's going to be larger, and actually what I mean here is not quite the average value of the F statistic. What I mean is the so-called "noncentrality parameter," but it's, that's a distinction that probably isn't important for the general point.

And, in fact, you can see that the ratio of the noncentrality parameters under the two models with due allowance for some little cheating on notation depends on this factor. It depends. Well, first of all, you can see that if all these numbers are bigger than 1 in this fraction, then, lo and behold, the fraction will be bigger than 1.

The crucial issue is if there is heterogeneity in the treatment effects across blocks, the fixed effects analysis is going to have, is going to have a larger noncentrality parameter. It's going to be more powerful.

This omega parameter here is going to crop up again. It's going to turn out that when we introduce blocking factors, and particularly if those blocking factors are not considered fixed effects, then the amount by which the treatment effect varies across blocks is going to enter into the sensitivity of the analysis.

But, remember, that shouldn't surprise you. If you think again about what this unconditional analysis is in this design, if you think about computing block specific treatment effects and then doing, asking whether the average of the block specific treatment effects is different than zero, the sort of intuition is, well, how am I going to do that?

Well, I'm going to take the average of the treatment, the block specific treatment effects, and I'm going to compare it to the standard deviation of the block treatment, at the block specific treatment effects, and so if there's heterogeneity there, it's going to be harder to detect effects.

You know, the details of the symbols don't matter, but this is sort of a general idea here that I think is accessible. I hope it's accessible.

Now, remember, the question that I was asked, essentially, can I kind of ignore the blocking, suggested that the statistical analysis was going to be not linked to the sampling model necessarily, and so there are three possible things you could imagine doing for the analysis of your now blocked design.

One is you could ignore the blocking all together, and that's what the impetus of the original question is: can't I do this and then not pay attention to it? You could include the blocks as fixed effects. You can include the blocks as random effects.

And the point is which of those is a good thing to do depends on the kind of inference you're hoping to make. If you're interested in making unconditional inferences, that is inferences to blocks, let's say school districts beyond the ones you observed, then, well it's going to turn out ignoring blocking is always a bad idea.

But ignoring blocking here is a particularly bad idea because it really can inflate the significance levels in ways that you would find surprising. Your .05 significance level, if you ignore the blocking and carry out a test at the .05 significance level, it's possible that 50 percent of the time, you will reject the null hypothesis by chance when the null hypothesis is true.

So the actual significance level might be 50 percent or even higher in plausible circumstances. So it's a really terrifically bad idea to ignore blocking in most cases.

Now, if you're interested in making unconditional inferences, including the blocks as fixed effects and doing that analysis is also a bad idea, and it also leads you to anti-conservative decisions about hypothesis tests for treatment effects.

Including the blocks as random effects, you know, is the right thing, but it's an analysis that has less power than the conditional analysis. Now, that doesn't mean it's wrong; it's an analysis that's making a different inference.

You ought to expect the analysis that is suited to the unconditional inference to be less sensitive than the analysis suited to the conditional inference. Well, why is that? Because it's real. It's a lot harder to figure out whether there's a treatment effect in a whole population of blocks, most of which you haven't observed, than it is to figure out if there's a treatment effect in a specific set of blocks which you have happened to observed all of.

Okay. If you want to make conditional inferences in this case, that is you really only care about inferring about treatment effects in the set of school districts you've observed, then it's still a bad idea to ignore blocking because that actually has the reverse effect. It actually can lead to deflating significance levels. In other words, you may have huge downward impacts on the power.

You can include blocks as fixed effects, which is the right thing to do. You can include blocks as random effects, but this may also have the effect of reducing power and making it very hard to detect the effects when they're there.

So the notion that it’s always the right thing to do, if in doubt, assume that the blocks are random effects, is not necessarily a good policy.

Now, if we move on to the designs—I’m going to move on to the cluster randomized design, the hierarchical design, and point out that the kind of approach that I just outlined can be followed there, too. You have to think about blocking in the cluster randomized design as adding another factor to the design. The inference model is going to determine what the most appropriate analysis is, and you can reason that either way.

You can either reason from analysis backwards inference model or the other way around, but one important thing is that introducing a relatively small number of blocks here, as well as in the completely randomized design, may leave you with very reduced uncertainty to a larger universe of blocks based on a statistical inference kind of, you know, sampling theory generalization.

So what do I mean by it all works the same way for cluster randomized designs? Well, if the original design was that we have clusters assigned to treatments, I tried to depict here a case in which we have, you know, clusters listed across the top and some of them are assigned to treatment and some of them are assigned to control.

But because clusters are assigned to one and only one treatment—you know, no cluster is assigned to both treatment and control—when we introduce a blocking factor, what happens is, at least ideally, we get a set of clusters within each block that are assigned to treatment, and another set of clusters within that block that are assigned to control, and so the dashes there are intended to indicate that those are, you know, those are non-assignments. In a sense, the first m clusters in the first block are assigned to treatment, and the second m clusters in the first block are assigned to control, et cetera.

So this is a knowable design. It’s a partially hierarchical design. In this design, we see that treatments are crossed with blocks, but clusters are nested within block treatment combinations, and this has at least one immediate implication for, you know, what the design can reveal to us, and that is that since each block has both treatments, we can actually estimate block specific treatment effects here.

We can’t estimate cluster specific treatment effects, but we can estimate block specific treatment effects.

So we might imagine applying this design in a case where the blocks are school districts and clusters are schools, and we have some schools in each district assigned to each treatment.

Okay. Well, the usual partitioning, the original partitioning, if we want to ask how this impacts the analysis, the original partitioning would just subdivide the total variation into part due to treatment, part due to clusters, and part due to within-cluster treatment combination, within, well, within clusters that are within treatments.

And there’s a sort of standard analysis which would tell us that the right test statistic compares the treatment variation to the variation due to clusters.

Now, when we introduce this blocking factor, we’ve sort of introduced the possibility of a bigger partitioning of a total variation in the experiment into part that’s due to treatment, part that’s due to blocks, part that’s due to the block treatment interaction. Remember, this is a case in which each block has both treatment and control, so we can estimate a treatment effect in each block, and therefore, there’s the possibility of heterogeneity of treatment effects across blocks.

Then we also have a part that’s variation due to variation between the clusters within block treatment combinations, and then we have a part that’s due to within clusters, and if we ask, well, how should we evaluate the variation due to treatment, how should we, what should we compare the signal to, well, it’s sort of not immediately obvious, and it depends on the inference model.

If we think of the blocks as being fixed, that is we’re interested in a conditional inference about the blocks—now in this case, I’m thinking that we’re imagining that we don’t have all—that either we don’t have or we don’t wish to think of having all the clusters, all the schools within a district in our experiment.

But we are, but we might be interested in either inferring to just the set of districts in the experiment or some bigger set of districts.

Well, if we’re interested in the conditional inference, that is an inference that’s formally to the districts we happen to observe in our experiment, then we get, you know, one test statistic. We compare the treatment versus the variation across clusters within block treatment combinations.

If we’re interested in inferring to a larger collection of blocks than the ones we've observed, to a larger universe, an unconditional inference, then, the appropriate test—well, there actually turns out there isn't an exact appropriate test and analysis of variance, but there's an approximate one, and that's slightly conservative, and that one would compare the treatment variation to the variation across blocks.

Now, the key point here is that the exactly appropriate thing to do isn't the same in these two cases. And as you'd expect when we're making a conditional inference, there's a lot more degrees of freedom for estimating the uncertainty of the treatment effect, the more degrees of freedom in the error term, than there is under our unconditional inference. And so that would lead you to say, okay, the conditional inference is probably going to be more powerful than the unconditional inference, and that shouldn't surprise us, again, for the same reason it wasn't surprising before.

The thing that's maybe a little less obvious is—than just the degrees of freedom difference—is there's a difference in the noncentrality parameters, a difference in the average size of the test statistic. And in general, the average size of the test statistic is going to be bigger when there's a treatment effect under the conditional model than under the unconditional model; and exactly how much bigger it's going to be depends on how heterogeneous the treatment effects are across blocks and exactly how heterogeneous the clusters are within blocks.

And so this term actually, this omega rho B term sort of is a term that describes the amount of heterogeneity across blocks in the treatment effect, and row C here is the interclass correlation which describes—across clusters—which describes the amount of variation that there is across clusters.

Now, as before, there are different possible analyses. We could ignore the blocking all together. We can include the blocks as fixed effects. We could include the blocks as random effects, and which is the right thing to do depends on how we want to think about the inference model we want to make, whether we want to make a conditional inference or an unconditional inference.

As before, ignoring blocks is a bad idea. It inflates significance levels of the test, and it's a big effect. Including the blocks as fixed effects is a bad idea if we want to make an unconditional inference because it will also lead us to erroneous significance levels. And including blocks as random effects is sort of the right thing to do if we want to make unconditional inferences.

And we have a parallel situation that if we want to make conditional inferences to the blocks in the experiment, ignoring blocking is still a bad idea. Including blocks as fixed effects is the right thing to do. In this case, strangely enough, including blocks as random effects doesn't have, it doesn't have a big effect on the significance test, which is not something you would have expected but turns out to be true.

Now, I should say something about these sort of general pieces of advice here. When I started this talk, I thought it would be a good idea to give you some formulas for these things, and then I realized that they aren't easy to interpret. I will say that this advice that I've given here about the completely randomized design is predicated on typical values of some of the parameters that affect things.

If you block on something that actually has absolutely no effect, it turns out that you can ignore the blocks and it won't hurt you. But you're probably not going to be blocking on things—I mean I hope you aren't choosing things that have absolutely no effect to block on. You know, it's not a good idea.

Okay. Now, one can do the same kind of analysis for a multi-center, for studies that start out being multi-center, randomized block studies. And essentially what winds up happening is that you, as before, introduce a new factor, a new blocking factor, and the way to sort out what goes on in the analysis is to just, you know, sort of do the standard kind of procedure to sort through what the appropriate test statistics are in the design that you've created.

Now, it's interesting that both in the case of the hierarchical design and in the design that's originally hierarchical, the cluster randomized design, and design that's originally randomized blocks study, you get different variance of partially hierarchical experimental designs. They are known designs.

We know how to analyze them and all this stuff, but they aren't exactly the same, and, you know, I have material on this, but I don't think I'm going to go through the details of what happens in the multi-center design when you add extra blocks because, in a way, the sort of message that came from the first two cases we've talked about probably you've gotten.

Okay. I got to figure out whether I want to make conditional inferences or unconditional inferences, and then I got to sort out the right thing to do. Make a statistical analysis that corresponds to that.

So what I'd actually like to do at this point, because this could actually, this could actually lead to our having a little bit of time for a discussion, and if I'm left to my own devices and I go through all of this, there will be no time.

So I'm going to speed through this, and get to "another easy question."

[Laughter.]

MR. HEDGES: And this is a question that comes up all the time. There was some attrition from my study after assignment. Does that cause a serious problem? And, you know, this is another simple question, and the answer is far from simple.

I have an answer, and I'm going to frame it in terms of experimental design, but before I do, let me just point out that there's a simple question that has an easy answer, a different simple question that has a really simple answer.

And that's 'does attrition cause a problem in principle?' And the answer to that is 'yes'. You know, randomized experiments with attrition no longer give model free, unbiased estimates of the causal effect of treatment.

Remember, the reason we love randomized experiments is because they in principle give us model free estimates of treatment effects, causal effects of treatments. And we lose that if there's attrition.

Whether the bias is serious or not depends on the model that generates the missing data. That’s an old refrain I’m sure you’ve heard before. But before we, before we go too far with this, let me just give you one way of thinking, thinking about post-assignment attrition in terms of concepts that you’re used to from thinking about experiments.

If we have a treatment and a control group, we can think about missing this as like introducing another factor. We have the observed data for the people in the treatment group and the control group, and then we have the missing data for the people in the treatment and the control group.

Now, by just making this picture, you can see that there’s a problem in estimating the treatment effect from only the observed part of the design. The observed treatment effect is only part of the total treatment effect. You could think in—another term sometimes used in experimental design is that the observed data allows you to estimate the simple treatment effect on the observed, but doesn’t allow you to estimate the treatment effect on the missing, and the main effect of treatment is a combination of those two.

But this is trickier. You know, there are things you can learn by just thinking about the algebra of this that are useful. Suppose, so now what I’m going to do now is talk about algebra, not talk about anything that has to do with statistical inference or anything deep.

This is, the rest of this bit is just about, just about algebra. If we think of our design, and now we’re taking the God’s-eye view of our design. We’re assuming we can see the population means, not only for the observed people, but for the unobserved people, and this helps us kind of understand what’s going on.

If mu TO is the mean in the treatment group for the observed individuals, and mu CO is the mean for the control group of the observed individuals, and we have mu TM is the mean of the missing folks in the treatment group, then mu CM is the mean for the control folks who are missing, and again you don’t have access to these quantities. You don’t have access to any of these quantities in a real experiment.

But for the purposes of understanding what post-assignment attrition might do, we can imagine we had access to these quantities. We could imagine ourselves in the role of the deity in actually understanding what’s going on beneath the sampling error.

Now, I’ll introduce another quantity which is obviously important. When that’s the proportion of the total number in each group that are observed or missing, so pi T is the proportion of the treatment group that’s observed, and pi bar-T is one minus pi T. So, in other words, it’s the proportion of the treatment group that’s missing, and the same thing for the control group. So pi C is the proportion of the control group that’s observed, and pi bar-C is the proportion of the control group that’s missing.

It’s clear that all together this is the information you would need to sort out the treatment group, the actual treatment effect on the total group that was randomized, and a little bit of algebra tells you that, well, pi T times mu TO plus pi bar-T times mu TM, that’s just the mean of all the people in the treatment group, both the observed and the unobserved.

And pi C mu CO, plus pi bar-C, mu CM is just the mean of all of the people who are in the control group. That is all of the people who are randomized, the missing and observed.

And so the difference between those two is just the treatment effect on all individuals randomized, and when the proportion of dropouts is equal, this simplifies further; so that if the proportion missing in both treatment and control group is the same, and I'll call that pi, then we can write the treatment, the treatment effect on everybody randomized, in terms of the treatment effect amongst the observed and the treatment effect amongst the missing in just the form of the last slide there.

In other words, we could write, if we called delta the treatment effect on all the individuals, then delta is just a linear combination of the treatment effect among the observed folks and the treatment effect upon the missing folks. And these two pieces are weighed proportionately, you know, their proportion pi of the folks that are observed.

So the treatment effect among the observed gets weighed pi, and pi bar of the folks are unobserved, so there the treatment effect among the missing gets the weight pi bar.

Now, this immediately tells us that we're not going to be able to do any inference on delta without making some assumptions or something that sneaks into the model, like assumptions, that constrains the possible values of delta m.

So the reason I said there was a simple answer to the question of whether or not attrition was a problem in principle is that the value of the treatment effect on the individuals randomized could be anything if there’s any missing data, and we have no knowledge of delta m.

So let me just sort of say that again because it’s an obvious but profound point. If we think of our outcome variable as having an infinite range or a practically infinite range, then I can make delta anything by varying delta m by enough, irrespective of what delta observed is.

So, in principle, attrition is an enormous problem that de-identifies our experimental estimates.

Now one way you can—now, of course, there’s a big, you know, caveat in what I just said, which was that if delta m has an infinite range, then dot-dot-dot, but we all know that the variables we measure usually don’t have an infinite range. You know, we have test scores that are bounded in some ways. We might even have some bounds we’d be willing to set on the treatment effect among the missing based on plausibility arguments, which would allow us to bound the total treatment effect based on knowing the pi’s.

So one way that people have approached the problem of dealing with attrition is to think about arriving at bounds on the things we don’t observe—this is sort of the Chuck Manski approach, you know, figure out bounds on the things you don’t observe—and if you can specify bounds, then you can specify bounds on the thing you want to estimate.

And so the key point is no estimate of the treatment effect is possible without an estimate of the treatment effect among the missing in this model, and you know, it’s possible that we can improve by assuming a range of plausible values. There are various ways to do this. One of the ways to do it is to take absolute bounds. The test scores go from zero to a hundred, then zero is the smallest they can be, and a hundred is the biggest they can be, and you could—well, I won’t go on yet—and you can set up bounds in that way.

You might say there’s a no qualitative interaction hypothesis, which is if the treatment effect is positive for somebody, it can’t be negative for anybody, and that sets up the possibility that if delta o is positive, then delta m can’t be any smaller than zero.

It’s an assumption, not data. It’s not guaranteed, but it’s an assumption that could be made and some individuals find that a very plausible assumption in, say, medical studies although you can easily think of examples in which it probably isn’t true.

But I want to go on for a minute because I think there’s another, again, simple point that is not, I think, completely understood by all scientists, and that is that when the attrition rate is not the same in the treatment and control groups, the analysis gets trickier. And one idea that people have used occasionally is to try to convince themselves about the treatment effect on those who drop out, those who are missing or unobserved, and they want to come up with bounds on that treatment effect.

And they say, well, you know, if I can be pretty sure about what the treatment effect would have been on the missing folks, then that will convince me that the attrition hasn’t spoiled the analysis.

Now how could you do that? Well, it’s maybe not so impossible to do. In a longitudinal study, for example, you may have multiple waves of measurements. And so you might use the treatment effect, the treatment effect measure on the last wave of observations you have, as kind of a proxy for the treatment effect later.

And/or you might have something you know to be strongly correlated with the outcome. It's not the outcome, but you believe you can predict the outcome from it, and if you predict the outcome from it, then you can get some kind of purchase on what the treatment effect upon the missing might be.

So if you got a situation like the one that I've depicted here, you might be fairly sanguine. So you look at this. I think this—it's made-up data, but it's the kind of data you might—it's not too crazy. We have a treatment effect among the observed.

We have an estimated treatment effect or a putative treatment effort, or maybe the deity whispered in our ear, and we know that this is the treatment effect among the missing. We see that the score, the average scores among the missing are smaller than the average scores among the observed, and that sort of makes sense, you know.

The people who can't make it to the post-test are sometimes not able to make it to lots of other things, and that may interfere with their overall achievement. But when we look at the treatment effect among the missing, it's just about the same—it's exactly the same as the treatment effect among the observed.

So how many of you are sanguine about the fact that the treatment effect among the total group of individuals randomized is going to turn out to be positive and maybe the same value as we've seen here?

I won’t ask you—you could raise your hands, but I won’t ask you to raise your hands—but I will ask you to make a mental commitment to that. Does it seem plausible to you that the treatment effect among everybody randomized is, yeah, going to be positive in 23 or something like that?

Pause. Do a little wait time. Let you think about that. Everybody convinced themselves they have an answer? All right. Well, it’s not, you know, obviously I set you up for this. I shouldn’t have set you up for it. I should have allowed you to say, to think what you might have thought.

But now I’m going to show you that just because the treatment effect among the missing is the same as the treatment effect among the observed doesn’t mean that either of those things is the treatment effect among the entire group randomized.

Here’s the rest of the data. Here are the numbers, and pay attention to the total on the far right. I rounded those to zero decimal places, but what you can see is that the treatment effect is positive and the same, exactly the same, for both the observed and the missing. And the treatment effect is exactly negative that for the entire group pooled.

Now, this may be a little surprising. It’s not surprising probably if you remember Simpson’s paradox because Simpson’s paradox, remember, usually isn’t—we usually don’t think about in terms of continuous data. We usually think about it in terms of discrete data, but this is just a simple example of Simpson’s paradox.

What’s going on? You know how can this be? The first time people see Simpson’s paradox, the question is always how can this be, and the answer to how this can be is that you see that most of the folks we observed in the treatment group, well, most of the folks in the treatment group didn't get observed, and the unobserved folks had lower scores.

Most of the folks in the control group did get observed, and most of the folks who were observed had higher scores, and even though there was this apparent perfect agreement between the observed and the unobserved in terms of treatment effect, the maldistribution of people in terms of observed and missing made it possible—and the difference between observed and missing on the average sort of made it possible for the average treatment effect among—the average score among the treated to be much lower than the average score among the controls.

This is just another example of, when I say Simpson's paradox, Simpson's paradox is why the death rate can be going up even though it's going down in every age category. How can that be? The population is getting older or the fact that the graduate school can be admitting more women, more of any category, in each department, but overall, have fewer of that group being admitted.

So the point is if you have unequal attrition rates, it's not enough just to know what the treatment effect would have been in the missing guys. You got to know something about what the treatment effect—you got to know the individual means.

Now, obviously, you can trade off information in coming up with clever bounds, and there's a whole literature on that stuff. One possibility for helping you understand the potential effects of attrition is to sort of put bounds on the treatment group means in some way, lower and upper bounds.

And if you do that, you can kind of reason through that if you put in the small—you put in the lower bound for the missing folks in the treatment group and the upper bound for the missing folks in the control group, you can come up with an absolute lower bound for the possible treatment effect.

And if you put in the upper bound for the missing treatment folks and a lower bound for the missing control folks, you can come up with an upper bound for the possible treatment effect on the entire population.

And in general there are many permutations of work like this. But I'll just point out that nothing that I said in this little section of the talk involves sampling or estimation error. It's all just algebra, and things get a little bit more complex if we start taking sampling into account, but really the biggest sort of stumbling blocks are the ones that I tried to identify here.

And now I'm at the stage where I'm going to start wrapping up so there will be a little time if people want to ask questions so I don't go for the whole time.

There are a lot of seemingly simple questions that arise in connection with field experiments, and I tried to cover one category of them and one other category to some degree.

And the main conclusion that I have about this is that the answers to these questions often are much trickier than they seem to be. They require complex thinking about fundamental aspects of what the research design is; what you're trying to infer from that design, and connected to it, what your vision of the sampling model is, even if it isn’t a formal sampling model in the sense that a survey sampler would recognize it.

And it depends critically also on assumptions about missing data and how seriously you take that. Now, if there’s only a tiny amount of missing data, obviously those problems aren’t so important. But if a third of your observations are missing, a third of the people disappear, then they are not to be ignored.

And I think the crucial point that I’d like to leave you with is that often these simple questions don’t have, you know, don’t have simple answers. They require a lot of complex thinking, and when you go to a methodological advisor, and they give you the answer that my friend Henry mentioned as the best answer to questions like this, “it all depends,” they’re not just trying to be uncooperative.

[Laughter.]

MR. HEDGES: Or it isn’t that they don’t know. They may not know, but, you know, a fair answer is ‘it all depends’; even then. I may know that much. So I think that the general sort of entreaty here is to be gentle with those of us who say I’ve got to know more. You need to explain more to me about what you’re doing and more to me about where you got your sample and what you think it means and what the design is because, for me, answering some simple questions, these and others, it’s hard to do without that additional detail and information.

I think I will stop, uncharacteristically, before the end of the time, and maybe there will be a possibility for discussion. I know there are several people in this room who know at least as much as I do about all these things, and if I don't know the answer or if I know the answer, but it's not right, one of them will help me out.

So I'll stop.

[Applause.]

MR. HEDGES: So are there questions or comments that people would like to make—maybe another simple question? That would be enough to fill up the time remaining, I'm sure.

I'm supposed to tell people to go to the microphone, and Henry has got first mover advantage.

MR. BRAUN: Thanks very much, Larry.

In trying to analyze observational studies, people often use selection models, sort of two-stage models, to try to capture the effects of selection in order to get unbiased treatment effects.

MR. HEDGES: Yeah.

MR. BRAUN: Do you know if people have tried to do similar things in the context of randomized experiments with non-random attrition to try to model the attrition in some way, maybe using covariates and so on, and have they been successful? Is there any way to tell?

MR. HEDGES: I'm glad it was an easy question. You know, there are different views about that. My colleague Tom Cook, for example, would argue that not so much of Heckman style selection models, but various kinds of matching models, you know, propensity score models and even just plain old garden-variety analysis of covariance can do a good job.

But I think the problem is that any demonstration that they can do a good job, and this is my—I know some of those folks are in this room so they may choose to get up and disagree and explain why I'm full of it. Demonstrations that they can do a good job in some circumstance always seemed to me to be—I have no idea how to generalize to another situation, and so I find them totally unsatisfying.

There's another point of view which is I know my former colleague Jim Heckman, when he was presented with some study—well, Robert LaLonde and Rebecca Maynard's work that did some selection model analyses of pretty good randomized experiments, and they got different answers. And Jim's response to that was, see, I told you experiments were unreliable.

[Laughter.]

MR. HEDGES: So it becomes, I think, so I'm not persuaded. What I'm persuaded of is that there are cases in which they can do a good job. What I can't tell is which cases they are, and that's the rub. I mean the reason I like experiments is because—and I think it's probably the reason all of us like experiments—you don't have to be very smart to do a good job with experiments. You don't have to know anything if you can keep the experiment intact. As soon as it falls apart, then you have to know stuff to make inferences. That's much harder.

Anybody from—Jack, okay. You're supposed to talk into the microphone. See, I remembered.

DR. RUBY: Name.

MR. MOSTOW: Oh, sorry. Jack Mostow, Carnegie Mellon.

When you are designing an experiment where you have multiple data points for individual, either single subject randomized within subject or randomized between subjects, is it exactly analogous to everything you said about clustering at the classroom level or do the rules change? And if so, how?

MR. HEDGES: It's almost exactly analogous. The only twist is—I mean this is actually one of the great things that Steve Raudenbush and Tony Bryk kind of contributed to our understanding of the world. I mean they didn't exactly invent this stuff, but they actually were the first—they actually told us about it, and you know, that's probably more important in some ways, that multiple observations within individuals are kind of a nesting structure that's very much like individuals within clusters.

And so much of what's true about, in fact, most of what's true, about clustering within schools or classrooms, is also true about clustering of observations within individuals. There's one sort of different point, which is that the correlation structure among observations within individuals can be a lot more complex than the correlation structure within clusters of individuals, at least the correlation structures we choose to model within clusters of individuals.

By that, I mean if you have multiple measures over time, it's highly plausible—well, let me back up and say the problem with, the problem that cluster—cluster randomized trials present us with is that individuals within clusters are correlated. They're correlated because if there's an effect of the cluster, everybody in the cluster has the same effect. They share it so their data is correlated.

But usually we take that correlation to be the same—that correlation among people to be the same. Now, when you model data across time within individuals, for example, it's possible that the correlation between measurements at any two time points is the same, but it's certainly plausible that measures that are closer together correlate more highly than measures that are further apart.

And the exact structure of that, and it turns out that the structure of that correlation matters, and so that's—the fact that the correlation structure is potentially more complicated is the principal way in which things differ.

And one could tear one's hair out about those things. That's, of course, what I've done, and—

[Laughter.]

MR. HEDGES: —if I hadn't encountered those problems, I'd have a full head of hair.

[Laughter.]

MR. HEDGES: So that's what I have to say about that. And again, I would encourage my methodologically-inclined colleagues I'm not the only source of knowledge in the room so—

MR. ALEXANDER: I've got a simple one for you.

MR. HEDGES: Uh-oh.

[Laughter.]

MR. ALEXANDER: Karl Alexander, Hopkins.

Indeed, a simple one. So I thought this was tremendously informative, very useful, even for someone who doesn't live and breathe this framework. So I am wondering if there's a way that we could get your complete list of simple questions.

[Laughter.]

MR. ALEXANDER: Can we write to you and ask for it? And your notes perhaps to walk us through some of the answers to these simple questions.

MR. HEDGES: Oh. Okay. Well, I could work on that. I don't know that I'll ever have a complete list.

MR. ALEXANDER: I don't mean complete, but a more complete.

MR. HEDGES: More complete. Yeah, I could probably, I could probably furnish a more complete list of simple questions with a partially complete list of answers, and I'd be interested to do that.

Actually, I'd also be interested in hearing from other people what their simple questions are, and actually, Karl, thank you for that, because it's, you know, anybody who has any of these questions is not the only person who has it, and I assume that I haven't been asked all possible simple questions. So, but amongst all of us, we could generate a good list of questions.

And in some cases, it requires a little bit of work to figure out, to think through the questions, and it's actually valuable. I find it valuable, but that's the kind of stuff I do.

Yeah.

MR. BORMAN: Hi. Geoffrey Borman, University of Wisconsin-Madison.

Maybe I'll have a simple question. One issue that comes up a lot related to your discussion about blocking—maybe not a lot, but it's happened to me more than once—is this issue when you're involved in a recruiting—schools, for instance—a group of schools may come forward, or you may actively go out and find a cluster of schools that are willing to participate in your experiment, and those schools maybe don't necessarily constitute a real cluster in the way that we think about them, in that they're not really related to one another. They may be, if it's a national study from all over the country, and really not have much in common at all.

What they do have in common, though, is that you were able to round up this group of schools at one time, and you know they're eager to know what their status is in your experiment, if they're treatment or control, and perhaps your implementation schedule depends on kind of this rolling cycle of randomization.

I'm just wondering how would you think about that generally and analytically where, you know, these clusters are more just built out of convenience rather than being associated analytically with the outcome measure, and how would you think about that?

MR. HEDGES: Well, I can tell you if it turns out they're not associated with the outcome measure, you're home free because they're not real clusters. I mean they're not clusters in the statistical sense.

The fact that they arise over time is in a way not so much of a problem, at least in my view. And I think this rolling recruitment is really more the rule than the exception in a lot of studies. I mean some people manage to recruit all at once, but it's something that is—I guess the crucial thing there is to sort of ask yourself the question—I would go back to kind of the inference model question:

You know am I happy to just infer to this group—and this group of sort of temporal clusters, if you will—and I might very well be—and treat these things as if they're just, you know, they're fixed effects if they have any effects and not worry about the heterogeneity of treatment effects.

Now, as soon as I say that, I think of one example in which it really mattered in the early schools where it had a much different treatment effect than the later schools, and then you were left with the puzzle about what to do. If you think in the sort of conditional inference sense, you could talk about the average effect of the early and the late.

I'm caricaturing your example a little bit, but it does correspond to one well-known example to some of us anyway. You know as soon as your—there is a sort of a price for, intellectual price for thinking in terms of fixed effects. You know you're talking about the average treatment effect in the clusters, in the blocks you happen to choose.

And what if the blocks have very different treatment effects? Well, the conventional answer is take the average in some way. Maybe you weight it; maybe you just take the unweighted average. And that's "the treatment effect."

But if "the treatment effect" is made up of very different individual effects, then it’s kind of unsatisfying. You know no effect plus really huge effect—no effect in one block, huge effect in another block, the average is somewhere in between, doesn’t seem like a very good description of the data overall.

So I think, I don’t know, from my point of view, Geoff, your question is really easy if there are no block treatment interactions in that case and really hard if there are. And if there are, then I think they’re going to be, there’s going to be an essentially contested question about what the right analysis is although one can describe, one can describe the data and say, look, the early adopters found a whopping treatment effect in this case; the late adopters didn’t find much. And there’s a real difference between the two.

And in the absence—and this kind of illustrates a weakness of experiments—if you don’t have anything else than just the experimental data to sort out the heterogeneity of those effects, then it’s really hard to interpret them.

And I think all of us who do this kind of work, even though we may be proselytizers for experiments at one level, are also proselytizers for collecting enough other kinds of data that you can interpret the experimental data. And I know I’m preaching to the choir in you, Geoff, because I know you do that routinely, but I think that’s probably essential.

Ah, now Vivian may very well want to take issue with what I said about, in answer to Henry’s question because we know she works in this field. So—

MS. WONG: Actually I was going to introduce, hopefully, maybe another simple question.

MR. HEDGES: Uh-oh.

MS. WONG: I was wondering—let's say in the case of a simple random assignment study, do you have comments or thoughts about doing the analysis with a regression-adjusted results or sort of simple treatment minus control difference?

I guess in some cases where I see that there's no obvious reason that randomization didn't work, that a lot of times people still present sort of these covariate adjusted results for experiments, and to me that seems like it could introduce more bias, but—

MR. HEDGES: Aah. Well, you know, there is work on that. There is Freedman's work on how adjusting for covariates can introduce bias in the randomized experiment. But it goes away asymptotically, and it's not very big. I guess one—I guess I would say I'm sort of more comfortable with centered analyses so that you're not adjusting the treatment effect, but you're just adjusting variation within treatment groups.

But I think if all you're doing is just increasing precision, I'm not unhappy with that, Freedman aside.

If the result is sensitive to that, then I'm a little more worried. Not just, you know, that p goes from one side of 05 to the other, but you actually get something that looks like a qualitatively different answer, then I'd be very worried about, but I wouldn't expect to get that in a randomized experiment.

So I do think it's—so I'm generally pretty comfortable with covariate adjustments because in the real world we need them to get enough power to do experiments with feasible numbers of schools. So I think if we decide we’re not going to do that, then we’re hoping IES gets a much bigger budget.

[Laughter.]

MR. HEDGES: And gives a lot of it to us. Okay. Go ahead, Carol.

MS. O’DONNELL: I won’t say who I am in case this is a stupid question.

[Laughter.]

MS. O’DONNELL: But I think it might build off of what she just asked, but how do you know, under what conditions do you use a variable as a blocking variable versus a covariate?

MR. HEDGES: You know, there’s an old and a new literature on this, you know. David Cox and Leonard Feldt and folks like that wrote about it. If you have a really high correlation, and basically it depends—the answer, one answer is that it depends on the form of the relationship between the covariate and the outcome, and if you’re convinced you know it, and you know it’s linear, or you can linearize it, if you can’t do that, then using it as a linear covariate is probably not possible.

If you do know that, then there’s still a question of whether blocking or analysis of covariance is more efficient, and it’s a little—that’s a little bit tricky, but I think the general answer is, and I think there’s at least one person in the room who might dispute this, I think the general answer is that if you have a really high correlation, the use of the linear covariate probably gives you a slight power advantage, and there’s a point at which blocking is competitive.

And so the sort of simple answer, one answer would be you try to figure out what the power would be under either circumstance and then pick the one that has the best power if you’re not worried about the potential bias introduced by using covariates.

And it wasn’t a stupid question. It’s a classic question, I should say, because I can think of three or four papers on it. It becomes tricky. But one of the things that I guess I should also mention is that it does—knowing questions about fixed versus random effects also can, you know, and conditional versus unconditional inferences, what are you inferring to can also enter into this. And I answered that kind of cavalierly not using my own rules.

You know if you think of the values you’ve observed, the values of the potential blocking variable you’ve observed as being sampled in some sense, then that has implications for the analysis of covariance because the classic analysis of covariance would argue that those things are fixed, and if you got them by sample, and you’re serious about that, and it’s one thing to take a sample and say, okay, this is the universe I care about, it’s another thing to say I took a sample and I think of it as allowing me to help generalize to a larger universe. So depending on how you think about the covariate, there may be implications for analysis.

Well, looks like I’ve exhausted you. Oh, no. Jack has got another question.

MR. MOSTOW: When you have multiple observations per individual—sorry—I'm still Jack Mostow—it hasn't changed.

[Laughter.]

MR. MOSTOW: But I might want to.

MR. HEDGES: It would have more interesting if it had changed.

MR. MOSTOW: When you have multiple observations per individual, you can control for individual differences by actually throwing in identity, student identity as a variable, and then that controls for everything whether you know what it was or not.

MR. HEDGES: Yeah.

MR. MOSTOW: But you may hear a giant sucking sound of all your power going away, or you can put in the covariates for the things that you suspect might be important but you might be missing something. How do you decide which?

MR. HEDGES: Well, now I'm getting into territory where I know people will disagree. In a sense, making, identifying, putting in a dummy variable for each individual is very much in the tradition of making the individuals fixed effects.

I mean that would be the language economists would use, and then the question becomes are you interested in this set of individuals and is between-individual variability to be—do you think—well, the question becomes is this a sample of individuals, and you want to generalize to a broader set of individuals or—and you want that to be part of the statistical rigmarole—or do you want to think of this set of individuals as being a set of individuals about whom I can learn things?

And if you want to think of the individuals as being, as being a sample of individuals that you hope your results generalize to, and you want statistical—you want the sampling theory generalization model to be the way to do it, then my answer would be you ought to treat those individuals as having random effects.

Now many—now I know a standard approach to this analysis is to make them all fixed—is to, in a sense, to make them fixed effects, but you don't have a sampling theory warrant any more for saying this should apply to other people from the universe from whom these people were sampled, and you have to make the extra-statistical kind of argument, a mechanistic kind of argument, for generalization.

Now that may not be; that may not be too big a price to pay because you get a lot by creating individuals as fixed effects. You get a lot of precision for estimating what's going on with them, but, of course, you don't get something for nothing. You got all that precision by defining away between-people variability as being fixed.

And it's the same, and it's not any different than I got a bunch of schools and, boy, there's a lot of between-school variability; maybe I should just, you know, make the schools fixed effects, and at that point, you get a lot more precision, but at the cost is that you've defined away a lot of the natural variability, and the statistics don't allow you to bring it back in.

Now, it may be that if you're trying to do something like a proof of concept, you probably want to do—it may be exactly the right thing to do.

If you want to show that something works someplace under good conditions, then the fixed effects approach seems like a good one to me. If you want to show that this is going to work if we scale it up, that sounds like a bad approach.

I'm expecting somebody to disagree with this, with that, but maybe I'll be lucky and we'll run out of time before they get a chance to calm down enough to say something.

Yes? No? Okay. Well, you know, I'm here for another three minutes so you can—

[Laughter.]

MR. HEDGES: You can ask questions if you like. Well, okay. I'll put in an advertisement then if we've got another few minutes. This is a great—the IES Research Conference is really a great gathering, and it is tremendously, well, gratifying to me to see how many good people there are doing serious education scientific work.

One of the other places where you can find lots of people who are interested in doing good scientific work in education is the Society for Research on Educational Effectiveness, a new scientific society that was formed a couple of years ago, and it was formed by people like you. Well, people like me in particular.

[Laughter.]

MR. HEDGES: And the idea was to have something kind of like the IES Research Conference, at least to have the same kind of population as the IES Research Conference. It’s open to anybody. I mean it’s not IES specific, but, you know, people who are interested in doing rigorous research on education, particularly research that tries to sort out causal effects.

We’re not a bunch of folks who just do experiments. There’s a lot of folks who do longitudinal studies, who do qualitative work, who do measurement, who do statistics, but the meetings—we’ve had a couple of national meetings so far. They’ve been a lot of fun, for me anyway, and, you know, I had to do a lot of administrative details so if they were fun for me, I think they were more fun for other folks.

I would urge you to think about joining the Society for Research on Educational Effectiveness or if not doing that, at least maybe you want to come to one of our conferences. The next conference is going to be roughly a year from now in early March of 2010. It will be in Washington. You can expect there to be a few hundred, a small few hundred, of individuals that are in fact a lot like you. In fact, some of them are you. But some of you probably aren’t part of it.

We have a journal, the *Journal for Research on Educational Effectiveness*, and I think one of the things that has been impressive to me about the group so far is that when I go to the conference, there are always more things that I want to go to than there’s, you know, copies of me to go to them.

And so I find that I want to go to the methodological talks, but I also want to go to a lot of substantive talks because they’re fascinating, and I also find that the people that I meet there are interesting and fun, and I learn from them, and I would commend it to all of you to at least think about.

You can find information about us at www.educationaleffectiveness.org, and or you can just sort of Google Society for Research on Educational Effectiveness. And anyway, I think that is another place where you can not only present your own work, but you can get a chance to see high quality, uniformly high quality educational research in a scientific mold, and it's a conference that's small enough that you can talk to people in the hall, and that's really part of what we try to do.

We try to create a program that gives opportunity for interaction, and we also do things like we provide food for people so that people don't run off at lunchtime and so they get a chance to stay together, and we usually have a few events that are yet another excuse for mingling.

So I would urge you to at least consider either joining the society or coming to our meeting or both and maybe even thinking about presenting your own work there.

We will look forward to seeing as many of you as we can. I guess it's okay for me to make that commercial.

[Laughter.]

MR. HEDGES: And if there is no other question, the fellow in the back was holding up a sign that was threatening to cut me off, so perhaps we should end.

[Applause.]

[Whereupon, at 2:45 p.m., the panel session concluded.]