Institute of Education Sciences

Fourth Annual IES Research Conference
Concurrent Panel Session

"Assessing Intervention Fidelity: Models, Methods, and Modes of Analysis"

Tuesday
June 9, 2009

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

Contents

Proceedings

DR. BUCKLEY: Good morning. I think we are going to go ahead and get started so if you can take a seat.

Welcome to the session on "Assessing Implementation Fidelity." My name is Jackie Buckley, and I'm a research scientist at IES in the National Center for Special Education Research, and as you can tell just from attending this conference, attending the sessions, seeing the posters, IES is certainly making progress in its mission of transforming education into an evidence-based endeavor that uses results from rigorous research to understand what works, for whom, and under what conditions.

I believe we certainly have made progress since the early 1980s when a certain David Cordray was involved in an article that recommended to Congress and the Department of Education that we needed to increase the rigor of education research and evaluation in this country, and you all are a testament to that, to that endeavor.

It is one thing, though, to employ a rigorous design in an education research study. It's another thing to truly understand the impacts of an intervention and understand what works and for whom and under what condition.

And doing that well, in part, means understanding important sources of variation affecting your outcomes and essentially affecting the utility of the research that you do.

And implementation, as you are well aware, is certainly an important source of variation that we need to understand, that we need to

measure, that we need to account for in the research that we do.

Historically, very few studies have published results of treatment fidelity, not just in education, but across various topics. Typically, what you'd see is at most a third of published intervention results actually reported on implementation fidelity.

We are making progress in that area as well. IES, as you know, in the Request for Applications, we force you to think about fidelity and how you're going to address fidelity in your research.

I also know from experience, however, in my own work, as well as working with many of you on your research projects, that assessing fidelity is incredibly difficult. Understanding what to assess, how to assess it, how to really truly take account of intervention fidelity in your analyses, and that's what your speakers today have been doing.

They've been figuring out how to do that well and help you figure out how to do that well. So I am pleased to welcome our speakers, actually I should say welcome our speakers back. If you were here last year, there was also an implementation fidelity presentation. So I am pleased that we are able to continue the discussion and further the discussion on implementation fidelity.

I would like to welcome Dr. David Cordray and Dr. Chris Hulleman to speak with you. I'll give you a little bit of background.

Dr. David Cordray is Professor of Public Policy, Professor of Psychology, Peabody College at Vanderbilt University, and Program Director for the Experimental Education Research Training Program, or the ExpERT

training program, which trains predoctoral and postdoctoral fellows in conducting experimental assessments to answer causal questions in education.

David's research is focused on estimating the numerical effects of intervention directed at at-risk populations. He has conducted multi-site evaluations of intervention programs and has greatly contributed to the development of methodological refinements of experimental, quasi-experimental designs; meta-analyses; and, of course, intervention fidelity assessments.

He is joined by Dr. Chris Hulleman, who is a perhaps soon to be or already former ExpERT fellow working with David, but soon he'll begin as an Assistant Professor in the Department of Psychology with a joint appointment as an Assessment Specialist in the Center for Assessment Research Studies at James Madison University.

Chris is a social scientist by training, interested in motivation and performance. He's currently involved in several projects that examine the impact of performance-based incentives on student, teacher and administrator motivation and performance. He's methodological interests include developing guidelines for translating laboratory research into the field and developing indices of implementation fidelity.

So we'll have about 25, 30 minutes for each speaker, and I ask that you hold your questions until the end so that we can get, they can get through all of their information, and we'll hopefully have a lively discussion for the last 30 minutes of the session.

There are microphones. This is being recorded so there are

microphones for folks to ask questions, and I just ask when you do come to the microphone, please introduce yourself so they have the information on the recording. And with that, I will give you Dr. David Cordray.

[Applause.]

DR. CORDRAY: Well, thank you. I forgot that it was 1980 when Bob Boruch and I did that RCT recommendation to Congress and the department, and so I don't feel so bad now that it took us that long to get to RCTs. So we started in 1980.

Some of what I'm going to talk about today is some material you've seen before, which is consistent with the 1980 to the 2002 time frame. So just to be warned. But what we've done a good job of making it better. So that's the consolation there.

The idea today is to talk about models, methods, and in particular some guidelines or some guidance as to modes of analysis regarding the incorporation of the rich data sets into the analysis itself.

I'm going to do the first part of this which really looks at the definitions, distinctions, and illustrations of fidelity, but also the idea of Achieved Relative Strength, which is something that we think ends up being critically important because fidelity cannot be done very easily in all instances because we don't have certain conditions filled.

I also want to put this in a context for RCTs. Fidelity analysis by itself can be done in a lot of different circumstances, but when you move to an RCT, there's a very specific set of circumstances and conditions that require us to think about fidelity differently.

And then more on the achieved relative strength as a special case in RCTs. I had hoped at this point that we would have a series of examples regarding modes of analysis where we could simply work through what it takes to do each of these kinds of analyses and what you get out of them, and the main problem there is that these things take a long time, and we're still in the field in two studies and the literature that we've looked at so far has not been helpful as part of our synthesis.

So what I'm going to be able to do is tell us about approaches that seem sensible as well as some of the challenges to those approaches, and then the last piece of this after I'm done is—no, not last piece—middle piece is Chris Hulleman is going to present a complete analysis that tries to follow as closely as possible the framework that we've laid out, and some of that is—that paper actually, the work was published recently, and it should serve as, if nothing else, an increase in his citation counts.

[Laughter.]

DR. CORDRAY: Which we hope that happens. And then the part that's most interesting about this is the kinds of discussion questions that come up and so we want to spend at least half of that time. So I'm to be pulled off of here at 30 minutes.

Some of the things that end up being important. We've got to distinguish what we mean by fidelity assessment and just regular old implementation stuff. They're related certainly, but they have different, different notions.

For the purposes of this presentation, I'm going to talk about

fidelity, which is in one sense at the other end of the extreme from just simple implementation analysis. The idea, though, is that at one extreme, and this is what the implementation world has looked like for many years, is we have a descriptive inquiry that focuses attention on answering questions that are really not guided by prior expectations but guided by good observation of what transpired while an intervention was being put in place.

And we've all seen these very nicely characterized studies that tell us what happened and not what should have happened. When you get to the fidelity side of the continuum, we're really talking about something that's based on an *a priori* model. So we have in our heads to begin some expectation about what should happen.

And then fidelity for our purposes is really the extent to which the treatment as it is realized, and I'm going to use these—I can't get too far from this—I'm going to use the small t with the superscript Tx to talk about the realized treatment, and the pre-stated intervention as a theoretical thing, and we'll talk about that as T superscript Tx.

All right. So throughout this, we'll make that distinction. The idea now is that rather than just describing what happened in the small t superscript Tx, we're actually going to look at the difference between what should have happened and what did happen.

So infidelity then is the extent to which the realized treatment differs from the theoretically specified one.

Now, you're all looking at me, you should be looking at me going, oh, that's not very realistic. How many times do we have theories that

are specific enough that would allow us to quantify what the value is for that treatment?

And we'll use a notion of strength in a moment, but I want you to remember that this is the extreme, and there are some circumstances like this, but mainly what we end up with is a picture of practice that involves a combination of some theory driven, some model expectations, but a lot of it is still descriptive in the sense of basically trying to specify what happened, what transpired.

So, aside from the extremes, we'd all agree that there's a descriptive side and there's this theoretical side; the sort of the pill notion of fidelity, the medical model of fidelity. Besides those extremes, there's not very much consensus in the field about what fidelity means.

It means all sorts of things depending upon who you talk to. One of the things that we have been doing, and you'll see at some of the poster sessions after this—I'll mention those in a minute—is we've been looking at the literature and trying to cull from the literature best practices as well as the notion of what fidelity means in the field across different subfields.

And what we end up with is basically the notion that there are three main definitions that are used. True fidelity is focused on adherence or compliance, and that is the extent to which program components are delivered, used, received, as it's been prescribed by the theory.

What distinguishes this from everything else in the world is that we have a stated criteria for success. Reading First was supposed to have 90 minute blocks of reading every day, and so in this instance, assessing the

fidelity with which local LEAs met that criteria is straightforward.

Did you do it for 90 minutes? Did you have a block set aside for 90 minutes or not? We end up with criteria. A lot of them are not that specific.

And one of the things that we find from looking at this broader literature is that this notion of fidelity is actually pretty rare. You don't find very many, even within studies that have them, very many criteria that are explicit enough that you could count the difference between what is found and what should have happened.

And I know you're grumbling, well, why are you doing this t, big T minus little t thing? We'll get to that. I don't know you're grumbling; I just suspect you are.

Second aspect of this is much more prevalent, and that just has to do with exposure. And we can talk about program intentions and not have to have any kind of criteria for success. What we really need to know is did the intervention expose people to the kind of components that are necessary according to some model?

And we don't have the idea now of being able to say how close was it? All we know is how much exposure was there: 53 hours, 47 hours of professional development—is that good; is that bad? 20 hours of professional development—good, bad? We don't know. All we know is that's the exposure level.

Now, we have that as the most prevalent notion in what counts as fidelity assessment, is just the sheer simple exposure thing. So you ought to

recognize right away that I'm going to be in trouble here to the extent that we can't make a distinction between treatment as theorized and treatment as realized if all we've got is exposure; right?

Well, it turns out that the third aspect of this that ends up being fundamental to RCTs is the idea that interventions can be differentiated. That is the treatment, the unique features of the intervention, are distinguishable from other things that appear in the control group or even other treatments, other models.

And this ends up having a unique application to RCTs because it follows the basic notions of what constitutes an effect. If the effect is the difference on average between conditions, we ought to be able to look at the difference in conditions on average and link those together. That we end up with a differentiated program, and we'll find out it doesn't matter whether you have a true fidelity index or whether you have an exposure index. This is the thing that saves the day.

So you guys could write a check to me every time you're able to do this, and I'll happily sign it over to my favorite charity, which is me.

[Laughter.]

DR. CORDRAY: See, I can goof around like that, but you can't.

Let me link this, then, to sort of notions of causal inquiry, and if anybody hasn't seen Rubin's causal model, you ought to. I mean it really is, it's really quite elegant and creates a foundation for a lot of interesting things. True effect under Rubin is basically the difference between conditions for the same person.

So if we really wanted to know what the causal effect of something was, we'd subject the same person to both conditions and just difference that.

That's a swell idea except it doesn't work. You can't be in two conditions at the same time, and so what happens is we end up with RCT methodology that just extends this to a group average difference between conditions rather than individuals. So now we have an intent-to-treat type model as an approximation for the true cause, causal model that we like, and that helps us greatly.

So now we've got as our effect, we've got basically the difference on average between conditions. I already gave this away, but it's not surprising that fidelity assessment, and I'm going to use fidelity broadly again, whether it's fidelity true or exposure, is basically, in RCTs the examination of the difference between causal components in the intervention and control conditions.

Okay. Now we've got those lined up. We've got the difference on average and now we've got the difference between conditions, and this is going to end up being more important when we start talking about, well, what is the cause of the average difference that we see in outcomes? It ends up wrapping itself around the idea that it's a difference in the conditions itself.

And what Chris and I have been doing is basically coming up with examples, as best we can, and some frameworks for—and some of the statistical properties of something that we'll call an Achieved Relative Strength. That's the index that tells us the dispersion between groups on

average and provides us with a way of indexing something that is analogous to an effect size for outcomes.

Chris is going to tell us more about those indices shortly, but the Achieved Relative Index is basically—Achieved Relative Strength Index is basically the treatment as realized minus the control as realized. Whatever that difference is, is the Achieved Relative Strength.

And the nice thing here again—that's the reason why you're going to send your checks—is that this is a default regardless of what kind of measurement you're using, whether it's an exposure index or a real fidelity index.

I just want to put this back into perspective of how to link these pieces together so we're clear. Let's suppose that we have an intervention that we're thinking about. We believe that the intervention is going, the t, the Y bar sub t, is going to push the outcome to about 90 points, whereas, what would have happened otherwise, the control condition, it's going to stay at 65. Okay. A 25 point difference. All right.

When we think about power, this is the first thing we're thinking about. We don't know it, but this is what we're thinking about. Or, I guess we do know it, but we've got into a noncentrality parameter, and that makes it a little less interesting, a little less visible.

Here our estimate for power would be an effect size and assuming full fidelity of .83 if that difference is 25 and a standard deviation pooled is 30. So our expectation is an effect size of .83 with this model. And we powered up for that.

What we haven't been as clear about is the simple notion that behind each of those averages is a notion of the treatment and, in particular, the strength of the treatment. Some treatments are big strong babies. Others are weak and don't have much of a difference between the earlier condition.

Some of those can be turbocharged with mechanisms. Others basically show no effect. But if we just for the moment take the idea that strength is a useful concept, even though we can't measure it at this point, we see that there's a connection between these two; right. In theory, our power analyses basically suggest to us that the difference between c and t is sufficient in strength to produce a difference in the outcome. That's what's behind the noncentrality parameter.

So what we expect in relative strength is 25 units. That's the difference between the strength of T superscript Tx and t superscript c.

In reality, we end up with a small t, superscript tx, and a small t for the control, and that's basically arguing that there's at least two models going on; a model for the control group and a model for the treatment group. There's some reason to believe that educational practices yield 65 points on the scale under old circumstances. There's a model behind that. It's not just random.

And our new model is that it produces a 90 point value under this new theory, recognizing that the theory in practice is not the same as the theory in theory.

[Laughter.]

DR. CORDRAY: You knew that was going to happen. We end up

having to account for that, and that ends up being two sources of infidelity. There's an infidelity that's associated with the departure from the true treatment, and there's an infidelity that is associated with departures from what the control conditions should have been, but things happen.

Whoops. Wrong way. The Achieved Relative Strength then is, in this case, it's 15 units, not 25, because we've come up on the control and come down on the intervention, which then means that our achieved effect would be a half a standard deviation unit down here, .5, rather than the expectation of .83, as a function of that reduction in the relative strength. Relative strength was big to begin with. It gets smaller as a function of infidelity, infidelity coming from two sources: reduction from treatment and an enhancement of the control.

So far so good? Why is this important? Good question. Thanks, Dave. Things don't get—well, we can put this back in the context of the Shadish, Cook and Campbell threats to validity, and it turns out that our big one, the one that is probably the thing that gets the rest of analysis started is being able to pass statistical conclusion validity. If we can't detect covariation, it's a little hard to make any claims about our causal inference if they don't co-vary. So we've got to make sure this one gets right.

Variations in participants' delivery, receipt of the causal variable, the treatment, increases error and also reduces the size of the effect, dropping our chances of detecting covariation, which we all will recognize minimizes power—reduces power, not minimizes it—reduces it; right?

If you don't think that is true, I've modeled this after one of the

projects that you should have data on, but it's not being cooperative. We expected the effect size in this to be about .3, and powered it appropriately with 30 units. For randomization, j equals 30. Intraclass correlation of modest means, .13, and with 30, with 30 cases, effect size of .3, our power is very good if we ---.

If we drop in fidelity to .8, the power drops to about .57, and if we drop to 60 percent of the original intervention, the power drops to .4. Better off flipping a coin at that point.

Now, you might say, well, that doesn't bother me. Let me just make the study bigger. Right. What's the cost then of making that study bigger? Again, if we basically respecify, in this instance, I respecified the size of the effect that we're trying to detect basically as a function of the noncentrality being reduced by the proportion or fraction of implementation accuracy, what we end up with here—this is a little off because the pictures got a little balled up trying to put them on the slide—but at full implementation, we're back at power equals .8. That 23 really should be closer to 30, and I apologize, it got goofed up.

If we then go to 60 percent or 40 percent—I'm sorry—80 percent, we'd need about 40 cases, not quite, not—that sounded [like] an awful increase. But if we go to 60 percent fidelity, that is, it's 60 percent of what should have been there, we end up with about 70 cases, studies, in order to come up with the same power.

So the fidelity does end up creating some grave difficulties for us. We can build our way out of it, design our way out of it, but in fact, it

does cost us. It costs us in terms of research dollars as well as tremendous amount of effort.

Okay. If that's not enough, we go back to Shadish, Cook and Campbell's threats to validity, the idea that what we put in place and what we test is not the same thing that we thought we were testing—we thought we were testing big T; now we're testing little t—leaves us with the question of what's the cause?

The cause is no longer the same thing as it was before. Even if we think of it as a difference in conditions, it's not the same thing because if it comes out as little t, how much of little t is there relative to big T? So the cause has now changed. That's a construct issue, construct validity of cause.

Poor implementation takes the essential elements, and they're incompletely implemented, driving the effect down. This can also happen—that's just the top piece. We can also contaminate the control by allowing the intervention to be a part of the control condition.

We avoid that with cluster randomization. We try to avoid it with cluster randomization. So the contamination due to proximity or propinquity is not really a big problem.

And the last part of this has to do with unexpected preexisting similarities between conditions. So we thought that the control really was sort of not so good, but in fact, when we get out there, we find that elements of the treatment are actually in the control condition.

All right. Each one of those things changes that difference between t superscript Tx and tc. So we don't know what the cause is unless

we measure this stuff. Again, Shadish, Cook and Campbell's threat to validity about external validity. We need to remember that our causal generalization is not about our theory; it's about what we achieved, and the difference needs to be known.

If we're going to start talking about practice, we're going to talk about giving people an understanding of what needs to be put in the field, we need to be able to specify the conditions under which we achieve the results, and they may be much less than what we found and what we have in theory. So, for generalization, we have to have a proper specification of what the cause is.

This gets kind of complicated when you end up with multiple components, and here's an example. This is again based on an example that would have been a real example had the data been available to us, and I apologize. So we'll use it as a hypothetical.

But let's suppose we have a three-component program that has professional development, it emphasizes assessment, and for the purpose of differentiated instruction. Okay. Those are the three components. And what we find in the theory [is] that we need six units of professional development, eight units of formative assessment, and ten units of differentiated instruction for it to be complete.

But, in practice, we end up with three units of professional development, six of assessment, and it's seven of differentiated instruction. Those are the—that's the source of infidelity for that treatment. Are you serious? Really?

[Laughter.]

DR. CORDRAY: Okay. I will have to talk faster. We told you this is going to take longer; right? Okay. Can I negotiate with you?

[Laughter.]

DR. CORDRAY: We got the other side of this, which is the bottom half of the picture, which is the control, and we had 2, 2.5, and 3 units for each of those components respectively in theory. With the augmentations, it comes back at 2.5, 3.5 and 4.

So now we've got all these components not being what they were supposed to be. That's one way to do this.

[Laughter.]

DR. CORDRAY: If we were to go through and at the end of this basically difference those conditions, what we find from our fidelity analysis, we end up with about a half a unit of professional development, about—instead of four, which is what the theory said, we end up with 2.5 units on assessment—something missing there—instead of 5.5, and we end up with 3 instead of 7 for differentiated instruction.

So that's now our new cause. Suppose these little puny differences actually made a difference? The cause is being created not by these whopping big changes that are expected, but these little ones, and that's actually good to know as well.

Chances of that happening are about as good as me getting through this in the next three minutes. Let me do this real quick. I have to do this real quick.

So use this as little review. True fidelity, we're up at the top up here. That's the only place where we see true fidelity. Anything beyond that is a difference between the conditions.

Exposure is also important, but doesn't say anything about fidelity. Okay. It's just this t superscript tx.

Contamination, augmentation of C or intervention exposure ends up getting us up above what we would have expected in theory by the control. Our saving grace on both of these is that we can think about this as treatment differentiation, intervention differentiation, and the one I like the best is the possibility of positive infidelity.

[Laughter.]

DR. CORDRAY: Okay.

AUDIENCE PARTICIPANT: Yeah.

DR. CORDRAY: Yeah. My wife looks at me every time I say that, and she says that can't be a good thing.

[Laughter.]

DR. CORDRAY: But in this instance, this context, this context only—

[Laughter.]

DR. CORDRAY: —positive infidelity is possible when people in the ground, the practitioners—this is going to be—never mind—the practitioners and people who actually know what they're doing are working beyond the constraints of the theory.

Now, we have, we have a project that involves tutoring, that it's

possible for the tutors to actually do better than the model said. They'd be infidels by our method. I hope that doesn't get me in trouble. But, in fact, they're doing a better job. They're doing more than what would have happened otherwise. So let's give credit.

Last time I said, oh, infidelity is bad. No, no, infidelity is good. As long as it's positive infidelity. So we can do that. So anyway, that was the reason why—one minute—not possible.

DR. HULLEMAN: Take five or ten more.

DR. CORDRAY: Really?

DR. HULLEMAN: Yeah, go ahead.

DR. CORDRAY: Okay. You're a blessed soul. That means you have to talk fast.

All right. Now, we get done with that part. Now we got to turn it into something that basically tells us about fidelity. How do you index that?

One way that's very popular in, not in education so much, but in other areas, is to essentially collapse everything into received or didn't receive, yes/no, dichotomize it, and then look at compliers, look at no-shows, and we can essentially take account of infidels, no-shows and the cross-overs, in the analysis. That's done a lot, and there are analyses for that. We'll come back to that.

Oh, wait a second. Stop. Structural flaws, and this is different. We started worrying about what happens, not just what happens in the pedagogy, but what happens in setting up the surroundings for the pedagogy. Incomplete resources and processes. Huge issue. External constraints. We've

seen things happen like snow days; you just can't get to class. It takes away from the amount of time and reduces the intervention. Strength for reasons we have no control over.

The more likely thing we think about is incomplete delivery on the part of individuals for core components. They just don't get it all done.

Now, here's a generic tutoring program. I won't tell you what it is for fear that they'll beat me up afterwards, but what we find in looking at one of the projects that we know about is that for this kind of tutoring, it's four to five sessions a week, 25 minutes per each session, 11 weeks, so our expectation of exposure is going to be between 44 and 55 sessions; right?

Now, the question is how are they doing? And so our first structural fidelity assessment is to basically look at what happens in this particular design which involves three cohorts, three cycles of kids. Kids get randomly assigned to the cycle, and the tutoring is delivered. The average number of sessions in cycle one is 48 sessions. Cycle two is 33 and cycle three is 32.

So right there, just structurally, before we ever get into the pedagogy, we end up with a structural infidelity that has to be accounted for.

And it's not just structural. Another aspect of this has to do with tutors, and this instance, this particular project, there are 18 tutors. And what we have in this slide is basically the average number of tutoring sessions per tutor, and just to be quick about it, if you look at tutor number two over here, about 25 hours on average and contrast to tutor 14, which is about 47 hours, kids in those two tutoring conditions will get quite different exposures to

tutoring, another issue of infidelity. Not just pedagogy but structural. But this actually had something to do with pedagogy I suppose.

The second part of this study is not to just look at these kinds of structural characteristics, these kinds of issues of individuals and their performance, but rather to look at the extent to which the tutor is consistent with the model, and we're currently working on an analysis or guys are working on an analysis of how faithful the individual tutor is to the model itself, which you think about that. It's actually sort of interesting because how do you get something that is completely individuated and come up with a fidelity score? Chuck Munter and Annie will tell you. It's fabulous stuff.

In practice, here's what we got to do. We've got to identify core components of these things. If we go off and just look at anything that comes up, anything that we can look at, we are going to be swamped.

Change models are terrific for this purpose. I am going to tell you about four sessions. Cat Darrow is number 31—after this— talks about some of these fidelity issues. Chris and his gang, 38, are talking about what we've done with these models of change. Michael Nelson is 44. He's talking more about models of change and logic models. And Kelly Puzio and Evan Summer, around here somewhere, 45, are talking about actually looking at training and the extent to which the training in this particular program was done as intended. Actually a very difficult thing to do. All splendid proposals.

'Establish benchmarks' is the second part, if possible. Measure those components, and using the logic models, and these guys are going to tell you about that as well, ends up being very helpful in telling us what to

measure, focusing us in on core components, and then what indices are actually driving the measurement indicators, and then converting them into some kind of a scale or an Achieved Relative Strength, and then incorporating them into the analysis.

Okay. Now, I'm about halfway through the presentation.

[Laughter.]

DR. CORDRAY: Do the essential stuff. But you also got to get stuff that's necessary to support the essential stuff and leave everything else alone. You need a model. It all starts with the model, and in this instance, this is Reading First out of the regressionist continuity study, and this is the extent to which they characterize all of Reading First starting from legislature; how the funding works.

The core components that are really critical in this one end up being the things that have to do with reading, have to do with using instructional materials that are research-based, getting professional development, using the stuff in class, assessing and diagnosing kids, and then having a classroom organization that supports that.

Each one of those very complicated things as part of that model end up requiring measurement, and in this instance, what we've done here is just basically took the simplest one possible, and said that if we go from constructs before that end up being important for Reading First are reading instruction, support for struggling readers, assessment, and professional development.

Each of those things by themselves are very complicated and have

sub-components, and you can see just in the reading instruction, there's three, and each of those has a facet. There's a piece of it that can be differentiated.

And in this case, we've got just two for instructional time, and each of those then needs an indicator.

So in terms of actually getting to the measurement stuff, we got to start with the model, articulate the components of each of the models in exquisite detail, as was done in Reading First, and in this instance, what they end up with is across the components a series of sub-components that each have a number of facets themselves. There's—you can split it one more time, in other words, and they end up with indicators for each of those to the point where, in essence, over, over all of the four components, the ten sub-components, the 41 facets, they end up with 170 bits of information about the implementation of Reading First.

Some results that you can get from that end up being presented here. This is not complete because I had to use data that was existing and not actually reanalyze things. But what we find is that the instructional time for Reading First is 101 minutes reported by teachers; 78 for non-Reading First.

Our Achieved Relative Strength index is .33. So there's some disparity between them. Support for reading instruction, 79 percent of the schools reported that—teachers report that versus 58 percent, and it goes on like that.

83 percent for support for struggling readers; 74 percent in the non-Reading First. And we have Achieved Relative Strength indices on each of those, and I will skip the U-3 [ph] thing for the moment because that has to

do with an index of how you characterize the dispersion based in standard deviation units.

Got to do a lot of scale construction. You got 170. You got very small number of cases, especially at level two. The analyses can be used descriptively. For explanatory analyses, also known as exploratory analyses by some parts of that—yes—lots of options.

Let me just get to the hierarchy here, and I'm going to turn it over to my friend. So you got all this data. What are you going to do with it? First things first. The intent-to-treat estimate that comes out of the experiment is—that's the thing that we want. That's the thing we prize. That's the one that we can trust. That's the causal estimate of the average effect.

If you add a true fidelity index to that— this is simply descriptively added—you might find that the effect size in the results is .5 and the fidelity estimate is 96 percent. You're done. That's it. There's nothing else to add. There's no variability in fidelity. Therefore, the effect that we have can be attributed to the difference, the implementation of the intervention.

If you have something that is not quite that specific, like these indices that don't involve a fidelity target, you just add the Achieved Relative Strength Index to that, and so Chris has got the study where the initial analysis shows that the effect size, intent-to-treat effect size is .45, and the Achieved Relative Strength is .92.

Again, we could argue a little bit and talk about how to structure

that, but the first things first is intent-to-treat and fidelity. They go hand in hand. They always ought to be there. You don't do an intent-to-treat without a fidelity analysis.

Next down the hierarchy that it is close to an experimental causal kind of claim is the Local Average Treatment Effect, which basically takes account of the fact that individuals are forced to be randomized, forced into treatment or not, and you could capitalize on that as a preliminary to an instrumental variables model where you simply adjust the ITT estimate by the receipt levels for treatment and control.

So LATE is our next one down. ITT plus fidelity maintains the causal analysis that we wanted. It maintains the period of the causal analysis. LATE approximates that.

And the things that we go from there, like treatment-on-treated end up being less understandable in terms of the causal effect of intervention.

The simplest one is an ITT estimate adjusted for no shows, compliance. A second one is to essentially use a two-level model, but basically come up with a modeling function for two, the second level where you have essentially an equation for control and an equation for treatment, and that provides us with a way of understanding the effects of implementation in the treatment condition, spectacular kind of analysis production function coming out of economics.

And then regression-based models that are more generic that exchange the implementation measure for the treatment exposure variables. We don't have zeros and ones anymore. We use the exposure variable.

Finally, we just do descriptive stuff. That's our lowest level in the hierarchy where we basically look at dose response. Good stuff. Why not do it if you can? Or partition into highs versus lows. I have a review that I did a couple years ago where we looked at ATOD prevention studies. If you look at the highs, the effects on average are about .13 to .18. When you look at lows, they're almost zero. So that's a nice demonstration that implementation fidelity matters.

Interventions are rarely very clear. These are the challenges. So it's hard to get this stuff. This is one of those things that I had hair like him when I started this 2 years ago.

[Laughter.]

DR. CORDRAY: You pull it out. So in 2 years, we're going to see what Chris looks like.

[Laughter.]

DR. CORDRAY: They are rarely very clear. So the models are very difficult. The measurements involve novel constructs. We don't have any history on these things. We don't know how to actually—we don't know whether the model that we've developed is psychometrically sound.

Fidelities at all levels, different levels. They are mainly used at second and third levels in HLM models, and you don't have that many degrees of freedom at second and third level. So you can't have 170 variables. You've got to do some fairly substantial data crunching construction.

Uncertainty about the psychometric properties, and we have a small number of degrees of freedom at the second and third level. You don't

have the kind of study size to go through and do a good IRT or factor analysis and all that stuff, and so you end up having to rely on good thinking about these things. Psychometrics are difficult.

And despite these challenges—oh, functional form. Boy, that's a real biggie. Don't know whether it's linear, nonlinear, asymptotic, and that can make huge differences in any of our estimation of the linkage between the fidelity measure and the outcome.

Now, despite all those challenges, the hairy guy over here, soon to be looking like Dave, has a dandy example for us.

[Applause.]

DR. HULLEMAN: Just talk fast. It's all right.

DR. BUCKLEY: I just want to note for our timekeeper in the back—timekeeper—we're going to try and hold Chris to about 20 minutes because I completely failed in holding David to 30. So—

[Laughter.]

DR. HULLEMAN: That's okay. I don't know that anyone could have. But I will try to be an example of tolerable adaptation and stick with the core components and discard everything else.

[Laughter.]

DR. HULLEMAN: And for people in the back, there are some seats scattered around up front if you wanted to find a spot. We're going to talk about here what I would call—and Dave likes to call—a simple intervention because it's social/psychological and I would say simpler. But the simpler, not necessarily theoretically, but simpler in the sense it has one

core component.

So it's not a realistic analog necessarily to most educational interventions, as we saw Reading First, multiple components. However, it allows us to exemplify the type of analytic process you could go through, and then you can sort of extrapolate that out for multiple components.

So we'll give that a shot here. So I'll outline briefly the theory of change here, which just says that we're looking at student interest and performance in classrooms, the two outcomes I think are really important. I put them on the same level actually. We can have that debate later.

But we designed an intervention that intended to promote student's relevancy and relevance in their coursework, and that was going to translate into their perceptions of usefulness of what they're studying.

So in this case, in this study, our dependent variable is going to be this mediator, which is their perceived value and usefulness, and we'll look at that through our study here.

So we actually have two studies. As Dave mentioned, these examples are outlined in the paper that was published in the *Journal of Research on Educational Effectiveness* earlier this year. So for the technical details, please consult that paper, and I will kind of buzz past through slides and won't even talk about those, but you can get those, and if you want to contact me, I've got the paper as well.

We have two studies here. One was in the laboratory and one was in the classroom. And they're designed to sort of replicate each other. One was extending the laboratory study into the high school classroom. Tried to

make as much similarities as we could, but you could certainly pick apart the differences. For example, in the laboratory, we taught students mental math. In the classroom, we're talking about ninth grade, mostly ninth grade science.

So slightly different subject areas, but mostly what's really important is the intervention was the same. Students wrote about what they were studying, and they tried to connect it to their lives. How was math? How was learning about the Krebs Cycle relevant to their lives?

And differences. We followed kids in the classroom for a semester. The lab session is an hour, et cetera.

So what did we get on the outcome? On the left-hand side, we have the laboratory study. Blue bars are control conditions; red bars are treatment. And what you can see on the left-hand side here in our motivational outcome is students' reports of how useful they find the course material or the laboratory task. What we see is we have an effect size of .45. Reasonable. Feel pretty happy about that in social psychology. Effect size of around .5, I feel very good about.

Move over to the classroom, all excited. We got this great intervention. Going to make a difference in students' lives. And we have an effect size of .05. Definitely nonsignificant. I could have gotten a lot more students and said it was significant. But nonsignificant. So that led us to say, well, what's going on here?

And as you might guess, we're going to talk about fidelity, and so because we had a single component, that is what students wrote about how well that this treatment helped them connect the material to their lives, we're

talking about exposure here, and a certain type of exposure, the quality of exposure, not just doing it, but how well did they make connections?

So what we did was we read all the students' essays. And we coded them for the extent to which they made connections. And so that is our single measure of exposure of fidelity in this study. And we had good agreement on inter-rater reliability, et cetera.

Now I want to just show you quickly the frequency chart. So the coders coded student writing from zero to three, zero being nothing, no connections at all, and three being very strong connections.

On the left-hand side, we have the laboratory; on the right-hand side, we have the classroom. And so you have your quality of responsiveness that the students wrote about, that exposure from zero to three, and you can see in the control condition, in the laboratory, we have no contamination. There is no augmentation. We have all of our laboratory control participants not making any connections.

And I should say the control conditions are just simply writing a summary of what they were learning about — so summary of learning or making connections.

In the treatment condition, you'll see we have a few students who weren't able to make any connections in the laboratory—seven were at a zero—but most of the rest of the students, 89 percent of them, were able to make some type of connection with over almost 60 percent making some strong connections.

The classroom, very different story. Again, in the control side, no

contamination or very little contamination. But you see we really have a dampening of the treatment effect of the quality of exposure in the treatment conditions here.

So just looking descriptively can tell you quite a lot. Now this doesn't mean that's why it didn't have an effect, but certainly we get the sense that there wasn't that quality of connection that we had hoped for.

So when we talk about indexing fidelity, as Dave has talked about, there's different ways that we can consider quantifying that. So one way to say absolute, that we expect everybody in the treatment condition to get a three. They should all get a three, or maybe it's a two, but we say, *a priori*, there's a level that we expect everyone to get, and we can measure that.

In this case, we scored people relative to three. There's nothing to say that that's right. I mean that's the coding scheme we came up with, but it might be we should have done it on a ten point scale and there was more out there that we should have expected.

So that's something that it's nice to talk about, but maybe it's difficult to actually pull off.

Average levels. This is simply the average levels of observed fidelity; in this case, our exposure index. And then we can come up with binary indices, and this is very common. You decide whether people actually receive the treatment or not. So we could go back to our frequency distribution, and we could say, okay, you have to score at least a two or greater to have us say that you received the treatment, and then we classify

people in the treatment condition based on that score.

And you can also classify people in the control condition, too, if you have a lot of contamination.

So I'm going to throw this up here. I'm not going to spend much time on it because Dave took my time.

[Laughter.]

DR. HULLEMAN: But also—

DR. CORDRAY: Just helping you out.

DR. HULLEMAN: Thank you. This is all in the paper as well, but what I'm trying to outline here is you have three different ways of conceptualizing fidelity: absolute; average; and binary. And at this point, there is no reason to say one should be primary over another—the three different methods that you could use.

I just want to show out sort of conceptually how you might calculate them. If we look at the absolute measure, and compare the laboratory to the classroom, you'll see the top line here, the very top line of our chart, in the laboratory, we have the mean X bar Tx is the mean of the treatment group on our fidelity variable, divided by capital T Tx, what's the absolute highest they could get? In this case, it was three.

So divide the mean, 1.73, by 3, and you get 58 percent. Compare that to the classroom, 25 percent. And you can go down. The story is similar on all three of these measures that the laboratory fidelity is much higher than the classroom fidelity. Again, consistent with our degradation in effect size of our outcome.

So if we want to index facility as Achieved Relative Strength, then we have to compare those two. We have to create an index that's going to summarize the difference between our treatment and control conditions.

So to do this in a way that we could quantify across measures and across studies, we just created an effect size analog. So we pool the standard deviation of our fidelity measure, whatever it is, and we create an effect size index we call the ARS, or Achieved Relative Strength Index, and in the paper we describe how we take—Larry Hedges has done some nice work with clustering with classrooms and how that impacts effect size estimates—and we throw that in, and that's the g. So Cohen's d is sort of your typical effect size estimate. And Hedges' g accounts for the clustering.

In the case of the laboratory here, there was no clustering, so g or d produces, if the clustering is zero, then g and d are equivalent. If you have higher amounts of clustering, it's going to impact the estimate slightly.

And in this, I guess I should say in this study, the interclass correlations this represents, the amount of clustering, very low, .01 to .08, depending on the variable. So we don't get a lot of adjustment, but we still did it anyway.

I'm just going to show you the average ARS index and how you would compute it. This is in the paper. Not going to spend a lot of time, but what I want to show you is that on the left-hand side, we have simply the group difference, the effect size estimate. Then in the middle we have a sample size adjustment, and on the outside we have our clustering adjustment; and you're simply starting with the means of your fidelity variable and

working towards there.

And I'm going to skip this. Okay. So, I have about ten minutes left; right. Okay. So to put this in, map this on to the chart that Dave showed earlier, on the right-hand side, we're going to, instead of the outcome, let's put fidelity. Let's put our fidelity score here. In this case, 0 to 3. And on the left-hand side, our hypothetical unknown treatment strength. We are going to assume that 3 in our fidelity score is 100 percent treatment. We don't really know if that's the case. We'll just play along here for fun.

Capital T, the theoretical treatment, should be at 3, and the control condition should be at 0. What happened when we went in and observed was that our mean levels in the classroom anyway were lower than this. Actually this would be the laboratory one. And then we had a little bit of contamination there, not much, but a little bit. And so those are both our infidelity. So degradation and contamination from the bottom.

Our Achieved Relative Strength in this case was 1.32, and how do we calculate that? Well, we took our difference in the classroom, which is the means of .74 in the treatment, minus .04 in the control condition, and we're going to use our formula, and I didn't put the rest of the g part in there. But essentially it works out to you pool the standard deviation, 1.32.

Now, at this point, we don't really know what size Achieved Relative Strength we need on our fidelity indices. Is 1.32 big or little? I mean certainly if you would use sort of Cohen's references, anything over one is large. But you would have to think that there's going to be some degradation, that it's not going to translate right to your outcome. If you have 1.32 on your

fidelity index, it's not necessarily going to translate directly to a 1.32 on your outcome measure.

So at this point, this is descriptive, and as more of us employ this technique, we can get a better sense of recommendations for how large you need to be able to have an effect.

And if we summarize our indices across, if we summarize the Achieved Relative Strength indices across the two studies, you'll see that basically what happens is we have a much larger Achieved Relative Strength in the laboratory than the classroom. It's around 1.0 difference.

So about one standard deviation difference in fidelity, and you might ask yourself does that matter? If we correlate then that difference in fidelity with actual outcomes, we see that it actually does make a difference, and as Dave outlined, this is less of a causal analysis.

But what we did is we said, okay, let's take people in the treatment condition and lab are people in the treatment condition who receive the treatment according to our binary treatment variable. And we said anyone who scored two or a three received the treatment, and so we look at the mean for those folks, standard error bars around that mean, and contrast that to the control condition, and what we see is that if you actually received the treatment, yeah, there is a difference there.

If we look at the classroom, it's a different story. The means are in the right direction, but the error bars are so much bigger. We see that this lack of fidelity really hurt us, and so on the right-hand side, we have—and as a matter of fact, we had, I think about 39 percent of the classroom students

were able to have received the treatment.

So 60 percent of our treatment participants did not receive the treatment. So we have obviously a problem with fidelity here.

And we want to ask ourselves, where does this come from? So this is [an] example of analysis. Once you've determined, okay, fidelity has made an impact, we have lack of fidelity, you can look at the sources of that, so moving forward consider this an efficacy trial. Moving forward then, how could we buffer that intervention and make it stronger so that more of the people receive the treatment?

And Dean Fixsen and his colleagues have done a lot of work on implementation drivers, things that aren't necessarily theoretically a part of the intervention, but might support the intervention so that it gets received and actually gets implemented.

So this would be a way to identify what are the factors, and we considered two factors: Could it be more due to the students' lack of responsiveness or was it a matter of the teachers not doing what they were supposed to do? Because in the classroom, the teachers were in charge of giving students the opportunity to write their essays and make connections, and so there's two possible sources of infidelity in this study, well, in the classroom study.

In the laboratory, it wasn't a problem, and we don't have those multiple levels. So what we have to do here is use a hierarchical analysis where on level one we model the treatment effect with the students, and level two we can model the teacher effects. And so based on this HLM analysis, we

can determine is it the teacher variables or the student variables?

And so the student variable we used was response frequency, how many times they actually wrote an essay, ranged from two to eight during the semester. Students might not write essay because they didn't want to write an essay. I was shocked to find out that some high school students go to class and they don't do any work.

[Laughter.]

DR. HULLEMAN: When I was in the classroom, it was much different than my high school experience. I thought, yeah, most of these people probably aren't going to go to graduate school. So it was a very interesting experience, but also students don't show up to class or they're not on time, they're unorganized. All those things could lead students to not respond.

But also they might not respond because the teacher forgot to give them the essay that week, forgot to have them do the interventions. So it's possible that teacher effects, and so you have the student response frequency and you have teacher dosage. How many opportunities did they give their classrooms to make connections?

And what we did, without talking in depth about the core component here, is what we found—and go to the paper or talk to me later about how we did the HLM analyses—but basically that baseline, you can look at the amount of variability at the student level and the teacher level, and then when you input your student response frequency and teacher dosage, does that reduce the amount of unexplained variance? And if either of those

variables do, you know, they're accounting for some of the infidelity, some of the lack of fidelity.

And what we found is that the student variable didn't account, didn't reduce. Less than one percent it reduced the amount of explained variability at the student level, but the teacher variable did. And so from this, we could say if we're going to move forward, we're going to say, okay, let's look at what we can do to support teachers in making sure they're more consistent across classrooms in providing opportunities to make connections, and that looks like it might make a difference in terms of the level of fidelity that we saw in the essays.

So to summarize, what do we find and why is this important for us? We found this degradation of effect size outcomes going from the laboratory to the classroom, and what we did is we took a look at three different ways to index fidelity: absolute, average, and binary.

And what we found is across all three ways, it was about one standard deviation difference in the amount of fidelity, higher in the laboratory than the classroom. And that these differences in Achieved Relative Strength did translate into differences in the outcome as we showed through that slide I showed with the binary treatment received.

And that the sources of infidelity were due primary to student or to teacher and not student factors, and so leading us to move forward in how we might support that and have more effective intervention in the future.

So having gone through that example, I want to point out just a couple key points and issues to summarize kind of what Dave and I have said

and what we're driving at here.

So, first, at a minimum, we have to identify our core components of the intervention, and in the case of the example, I just said it was pretty easy. However, if you talk to developers or practitioners, and you ask them what are the most important things you're doing to teach reading, you're more likely to get this long laundry list of things that they all feel [are] important in the classroom.

And it's hard work to go from there to, as Dave did, let's narrow it down to these five components or four components, but that's really key because you saw what happened. He had five components for Reading First, and it blew up to 170 indicators. So if you start out with 20 components, you know, multiply that out, you have way more indicators than maybe you want. Not to say that you can't have a 20-component indicator, but there's that process, that winnowing and narrowing process, that's really key, and that requires researchers and practitioners and developers working together, which is my next point.

And as outlined by Dr. Easton in his talk on yesterday morning, that we do need to involve the practitioners in this process, and the people who are close to the ground, and the people who have developed these interventions. We just can't simply come in as researchers and feel like, oh, I can determine what the core components of this reading intervention are.

We need to listen to what people are saying because it's not exactly clear. And oftentimes it's not specified, *a priori*, 30 minutes is good; 40 minutes is better; ten minutes I'm not real happy, but I'm glad you did

something. Doesn't really help us say, well, what's our benchmark. And those are conversations that happen over time.

So we need to develop the intervention models and the core components, and importantly, benchmarks, and sometimes this is an iterative process with efficacy trials going back and forth.

And as I've tried to demonstrate today, you can have tolerable adaptation, reduce some of the essential but not unique components to your presentation or to the intervention. So you're focusing on what is really crucial.

And I've got five minutes left, and I have one slide, so we're doing good. The other thing that's really important about the work that Dave started a long time ago, and I've just sort of piggybacked on here recently, is this idea that the causal, the cause and what is producing the change, once you implement it in the field, it's most likely different from what you thought theoretically.

And this isn't just important from an academic perspective because you might, it's easy to think, well, the theory is important, you know, whether Professor X is more right than Professor Y, and whether or not you get tenure because you're published.

Sure, that's one way to look at it, but more importantly, I think, broadly, not that it's not important for those researchers to get tenure, as someone who is going to start this process, but more important, broadly, is this idea that we need to know what it was that made the difference so that, one, we can import that to new contacts, and we know what's tolerable to

adapt in situations and what we want to be rigid on and say you need to do this.

It's the benchmark really is 45 minutes of reading, and if you don't have that, it's just not going to work. So to know what the causal elements are help us communicate to practitioners what are the things that just don't mess with this, and other things that maybe are less crucial, as well as improving the intervention in the future so we can move forward.

And, of course, once you have identified those essential causal pieces, then you can see what varies with your outcome, and you're in a position, not a causal position necessarily, but from a correlational aspect, to really figure out what's associated with the outcome.

And finally, I guess the summary point is this is sort of a post experimental respecification. So if you consider, you know, theories are not static and they're changing, interventions may not be static and may be changing, and so from a continuous quality improvement perspective, but also from the idea of what's making the difference and what's driving what's happening.

So thank you, and I guess we'd like to open up for general questions now.

[Applause.]

DR. BUCKLEY: And just a reminder, if you do have questions, please use the microphone and introduce yourself before you ask your question.

DR. MAY: Hi. I'm Henry May, University of Pennsylvania.

Given that there's been a lot of work by Rubin, Bloom, others, in estimating causal effects under less than perfect conditions in experiments, in the case of where there are no-shows or cross-overs, we often take intend-to-treat effects and adjust them to produce effects of treatment on the treated.

I could see how you could actually do the same sort of thing with the fidelity measure. You could actually adjust your intent-to-treat effect to produce an estimate that showed what might the treatment effect have looked like if it had been implemented faithfully.

What are your thoughts on that?

DR. CORDRAY: Well, let me just do this real quick.

DR. HULLEMAN: Go ahead.

DR. CORDRAY: Part of the issue is the specific form of that model you're talking about, which is a LATE, Local Average Treatment Effect. If we try to think of it as a dichotomy for very complicated programs, it's not very well suited to that or you end up having to specify a level at which you think they get some, but that's not enough. We don't think that's enough. So the idea of fulfilling that particular binary model I think is difficult.

On the other hand, the expansion of the binary version of LATE to an instrumental variables model holds a lot more promise, I think. You still have issues of how much it's going to be enough to qualify as a complier or a defier, but—not a defier—as a complier or cross-over, but it can work in our favor in multi-component kinds of interventions. But I mean it's inherently limited to the idea that it's a simple kind of intervention that

people do or don't take. These things are not simple.

DR. WOLF: Patrick Wolf, University of Arkansas.

I wonder if there's any place in your conceptualization for naturally occurring exposure of the control group to the causal elements of the treatment? And let me give you an example from a study near and dear to my own heart, the evaluation of the DC Opportunity Scholarship Voucher Program.

So there's random assignment. Treatment group gets a voucher that they can use to enroll in a private school of their parents' choosing. Control group does not get the voucher. A year later, 14 percent of the control group is in private schools. They were all in public schools at the start.

We're able to determine that two percent of them were enrolled in private schools, were accepted in private schools because they had a sibling who had a voucher, and so the school said, well, we'll take them both for one voucher. So we defined that as program-induced cross-over or contamination and adjust for it.

The other 12 percent gained access to private schooling outside of the voucher program. We considered that to be a naturally occurring part of the counterfactual and do not adjust for it in our standard impact estimate. So would you say that we're missing the boat on that or is there some place in your conceptualization for sort of that naturally occurring exposure?

DR. HULLEMAN: Well, I guess I would say in our model, then, you would include that as part of your assessment, incorporation of your

definition of fidelity, and so if you're capturing that, then somehow you could model that in one of the later, the lower, I guess lower on the hierarchic analyses. Wouldn't see—

DR. CORDRAY: In the middle.

DR. HULLEMAN: Yeah.

DR. CORDRAY: Let me just do this real quick. The point you're talking about is basically something that's ubiquitous in control conditions, is that what we thought was unique to the intervention actually ends up being a part of the control condition, and it's hard to call those cross-overs.

So the idea then is, I think, to split it up the way you have, which is you've got some legitimate cross-overs. What value that estimate has I'm not sure if you think of it is a LATE.

But if you think of it as intent-to-treatment on treated, where you model that explicitly, where you model, now you're not talking about the difference between zero and one, you're talking about on critical components, what level of achieved difference is there or, put it another way, for a more generic case: what difference does level, regardless of condition, make on each of the variables, make towards the outcome itself.

That's modelable, and it's telling us about—it's not a LATE; it's not a causal analysis. It's somewhere in between, but it's telling us something about exposure to causal variables but not necessarily to the treatment itself. So you're on the right track. Yeah.

MS. EDMUNDS: Hi. My name is Julie Edmunds and I'm from SERVE Center at the University of North Carolina at Greensboro, and we

have an experimental study looking at an entire school. So needless to say it's a little bit complicated of an intervention. We've got—we've identified the core models and the indicators and we're collecting data on those indicators. So now we're at a point of, okay, now what the hell do we do with all this stuff?

And so what I'm wondering is if you have any guidance around if you're collecting multiple sources of data, so we have data on the same sets of indicators from teachers, from students, and then from observation, school site visit kind of things. So if you have suggestions around synthesizing that, and then what to do if they don't agree?

So if the teachers are saying sort of one thing or have one level, one perception of the level of implementation and students have a different perception, for example?

DR. CORDRAY: How many schools do you have?

MS. EDMUNDS: We'll probably end up with a total of 20 with good enough data to do something with.

DR. CORDRAY: Okay. So psychometrics, IRT, things of that sort, factor analysis, are probably out. So you're really, I mean one of the things that comes to mind is a rubric of some sort that you use to put those three sources together, and then if they don't weight, if they don't say the same thing, if they're not convergent, you might end up with three rubices, if that's such a word.

[Laughter.]

MS. EDMUNDS: So treating them entirely separately?

DR. CORDRAY: Yeah. And you—

MS. EDMUNDS: Oh, okay.

DR. CORDRAY: I mean that's the only way I could think of. Right now you're really sort of, you're back to a characterization of the units, and you've only got 20 of them.

MS. EDMUNDS: Right. And the random assignment occurred at the student level. So we have sufficient power from there, and we have some student level measures also. So—

DR. CORDRAY: So some of the—well, then you're back to being able to at least do a good job, a better job, and allow correlation matrices to help you refine the measures themselves either through IRT or factor analysis or some such thing for the student level, but at the level where implementation is probably critical, school level, you really, you got 20 quantitative case studies.

MS. EDMUNDS: Right.

DR. CORDRAY: And Robert Yin has a very, very clever scheme on use of quantitative case studies of that sort, and it's perfectly applicable to this situation.

Go ahead.

DR. HULLEMAN: Oh, no. I was just going to add it seems like there's two issues. One is if it's really important to you to determine exactly what happened, who's right or wrong, and that's sort of, you know, a separate issue.

And there is some, there's becoming more work on that. There are

also in personality psychology, there's a lot of work on inter, different sources of observation and accuracy, and I could talk to you about some of those references after.

But the other question then, as Dave was saying, I think the other issue is you're using those as indicators of fidelity. So whether or not their mean levels are exactly correct is maybe less important as the variability. So that you can utilize that variability in your assessment in your analyses. So—

DR. CORDRAY: But with 20 cases, the analysis is going to be fairly limited.

MS. EDMUNDS: Thank you.

MS. DOLFIN: Sarah Dolfin from Mathematica Policy Research.

I had a comment and a question about fidelity analysis in an RCT context. So, first, to the extent that infidelity is due to the result of participant choice, we may have a selection problem in addition to a fidelity problem. So if a teacher offers more tutoring to a student who's struggling, then it's not just the fidelity issue; it's also there's a selection issue.

In some cases, we might consider this to be an outcome in the sense that, you know, suppose the treatment is the offer of a program with certain components to teachers. The extent to which they participate and take up these various components might be considered an outcome. You know what happens when you offer teachers this type of program.

So I guess I was wondering what your thoughts are about sort of fidelity analysis in this kind of context?

DR. CORDRAY: You have two, I think, different pieces here.

The first one is basically given the opportunity to be assigned or not, and the outcomes for teachers is really an experiment, a first order experiment, and there is no selection bias there that is consistent with the first part of your question, which is that you've got individuals acting on the part of the students either to add more or to take away, and that is a selection bias.

That is post assignment selectivity bias that is the main reason why my lowest level of, in the hierarchy of how to analyze these things, really is descriptive. It really isn't causal. It really ends up being you have to couch it very cautiously as a claim about the value of the intervention because it is a selection bias, and I'm not even sure an instrumental variables analysis would handle anything of that sort.

A mediation model might be even nicer. As long as you're into a correlational analysis, why not make it intervention, change in teacher behavior and change in student practice or change in student outcomes, because you already bought the farm on that, you know, so farms, so to speak.

No, farms are good things for people that are farmers and stuff. They're experimental. Agriculture and all that. I get myself in so much trouble.

[Laughter.]

DR. SHERIDAN: Hi. I'm Sue Sheridan, University of Nebraska at Lincoln.

And in many of our intervention studies, we are collecting data not only on exposure or adherence, but also quite a bit of quality data, so a lot of coding on the quality with which teachers are implementing

interventions or coaches are coaching their trainees.

And I wondered if you had any comments or thoughts on how we might begin to combine some of these different dimensions of fidelity into some kind of a metric or, you know, the best way to really handle the complexity of the construct of fidelity as we're trying to understand its effects on the intervention outcomes?

DR. CORDRAY: Do you have a benchmark that allows you to say enough is enough?

DR. SHERIDAN: Not enough is enough. We have in some of our studies—depending on which intervention study we're talking about now—we have a pretty good indication of what we consider to be the gold standard implementation levels, but in some we're still sort of exploring because they're developed, newly developed interventions.

DR. CORDRAY: Sure. So at the root then, you're basically in the exposure camp on fidelity, and I can't imagine why you wouldn't be able to take an exposure measure and a quality measure and basically create whatever index is necessary to capture the two aspects. High quality and high exposure has got to be better than low quality and high exposure. So—

DR. SHERIDAN: We might have some kind of a four cell kind of thing?

DR. CORDRAY: Yeah.

DR. SHERIDAN: We're looking at just a product of sorts.

DR. CORDRAY: That's what I would imagine, yeah. I mean it would imagine that quality is related—in quality, high volume of instruction

that's good—

DR. SHERIDAN: Yes.

DR. CORDRAY: —in my experience has been better than high volume that's poor.

DR. SHERIDAN: That's very bad.

DR. CORDRAY: Right. It's like when you take a long time to do a presentation and run your colleagues off the clock, that's like a zero on quality, you know, but lots on exposure.

[Laughter.]

DR. SHERIDAN: So taking both into account. Thank you.

DR. HULLEMAN: I was just going to add that one of the things that I try to help orientate people to thinking about, when we're talking about fidelity measures, we're talking about, these are psychometric scales, so you think about how you approach your outcomes is not necessarily so different.

So you have theoretical ideas about, you know, which of these scales should represent which constructs, and you can also take the empirical approach which things group together, and not necessarily that either of those is the right way. I think that's where this is a developing field and we're developing our understanding.

So you can think about those two approaches, but remember that neither one is right, but that these are psychometric scales just like your outcome variables.

DR. BUCKLEY: I think we have time for two more questions, and then we're about out of time.

DR. GREENWOOD: Thank you. Charlie Greenwood, University of Kansas.

I'm interested in your thoughts about the situation where fidelity would be measured repeatedly within a context of a study. For one reason, it may be of interest to see what is the time to full implementation; when do we go from very low strength to a high strength situation?

So if we were multiply measuring within a school year, for example, how would we model that within your model? Any thoughts would be appreciated.

DR. CORDRAY: Is the intention to use that index, that time ordered index in a level two model to predict changes in the—

DR. GREENWOOD: It could be a time-varying variable in that sense.

DR. CORDRAY: Because creating just a slope on each, for each case, would be a useful index, I would imagine, so that people that—it doesn't tell you whether they get to the level, but you could actually probably dummy code that as, you know, exceeds the criteria or not, but then also how fast they get to that level, and where they started from. I imagine you could collapse that multiple measurement into a slope of some sort.

DR. GREENWOOD: Okay.

DR. CORDRAY: An intercept.

DR. GREENWOOD: Thanks.

DR. CORDRAY: Yeah.

MR. GLOVER: Hi. Todd Glover, University of Nebraska.

I was curious, you talked about taking kind of components and breaking them down into indicators. In some cases, some of those indicators seem like they may be more important than others. Just curious about your thoughts on weighting various indicators and where we are with respect to doing that in the field?

DR. CORDRAY: Yeah.

[Laughter.]

DR. CORDRAY: It is the case that there's going to be some important weighting function. Not everything is equal, and I think that really is where you end up having to collaborate with the practitioners and ask Easton to give us some money to do that. Collaborate with the practitioners and the weighting function on those things is very difficult because you're really not—I don't, I could do it, you know, but it would be wrong.

So, but I think it really is a matter of us working together on those things and Easton is probably one of the collaborators we'd like to have on that. He's not in the audience here; is he?

DR. HULLEMAN: And that was the part of the presentation I managed to get Dave to take out. So—

DR. CORDRAY: That's right.

[Laughter.]

DR. CORDRAY: What else we got? You've been very patient.

DR. BUCKLEY: Yeah. I think we're about out of time so I want to thank one more time, David and Chris.

[Applause.]

DR. CORDRAY: Thank you so much.

DR. HULLEMAN: Thanks.

DR. CORDRAY: Don't forget to call. Call home. And those checks. Keep those checks coming in. Thank you very much.

[Whereupon, at 10:35 a.m., the panel session concluded.]