Institute of Education Sciences

Fourth Annual IES Research Conference
Concurrent Panel Session

"Revision to Mean, or Does Dosage Matter?"

Monday
June 8, 2009

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

Contents

Proceedings

DR. HOLLISTER: All right. Let’s go. I want to first say that in the—I guess Mark tells me that in the notes about revisions of panels at the beginning of your book, it says—

DR. DYNARSKI: It’s the online book.

DR. HOLLISTER: It was online. It said that Fred Doolittle wasn’t going to be able to give his talk because it hadn’t been cleared yet or something like that, and I think he was just living up to his name, you know, “‘Do-little’, if you can.”

[Laughter.]

DR. HOLLISTER: So, but then when they knocked Fred out, they just, instead of moving Mark and the title of his talk up, they just moved Mark up, and so Mark is going to talk about afterschool Programs even though he hasn’t done anything on them.

[Laughter.]

DR. HOLLISTER: But as he runs the What Works Clearinghouse, he knows, he can talk about anything. He knows the answer on all of these things.

Well, when they first called me up and they said on the phone we want you to be the moderator on the Dotage Panel. Dotage. Oh, my God. So I went and I looked it up, and it says a state of senile decay—

[Laughter.]

DR. HOLLISTER: —marked by decline of mental poise and alertness, and I guess I have to admit that I fit that. I qualify under those criteria. So I called back, and I said, well, I guess I'll do it. And they said, no, no, not "dotage." That shows you're in your dotage. Dosage. And so then I thought about dosage, drugs. Because in supported work, we worked with ex-addicts so I read a lot of this drug literature and got to know people at NIDA, National Institute on Drug Abuse.

That's like IES for scientific study. So NIDA promotes drug abuse, I think, just like the IES promotes scientific studies in this thing.

But then I thought, you know, well, what they're going to talk about, and also I'm on the Technical Advisory Group to one of IES's studies that has to do with mandatory drug testing in schools. So I figured that's what it was.

And I thought about overdose, and so I looked that up, and it said ingestion of a drug in great quantities, in quantities greater than are recommended or generally practiced. An overdose is widely considered harmful and dangerous and it can result in death.

Randomized control trials at the RCTs for IES. That's clearly an overdose problem in this thing.

Anyway, that will, I think, cheer people here to hear that we're going to talk about overdosing on randomized control trials. But I also want to recommend to you, before I let the other people talk, a couple of papers, one by Angus Deaton and the other by Guido Imbens, that appear in the National Bureau of Economic Research.

And Deaton calls those of us who advocate RCTs "randomistas," which comes from his work in Nicaragua with the Sandinistas.

[Laughter.]

DR. HOLLISTER: And so he’s sort of saying and people who don’t like randomized control trials, you can go to—here’s a big name in econometrics who’s saying, you know, ‘they don’t give you any useful answers.’

And luckily, Guido Imbens comes back with a paper titled—because he’s talking about instrumental variables as well—“Better LATE Than Nothing.” Local average treatment effect is LATE. So I recommend those to you.

What are we thinking about when we think about dosage? We’re going to come back to that a little bit later. In fact, here’s the structure of what we’re going to do. They’re going to make each of their presentations. Then we’ll have a little time for questions on the substance of their presentations, and then we’ve got a list of questions, methodological kind of questions, that come from this dosage consideration that we’ll go if time allows.

Okay. So what are we thinking about in terms of dosage? One of the things that came out of previous literature in other fields is whether the effect declines over time or grows over time. So trying to get the time path would be one aspect for looking at these studies.

And this was a case in both employment studies and welfare studies that we saw these fade-out effects and recovery effects in some cases. And cost/benefit analysis, we want to know how fast, because we have to project beyond the observation period. How fast does any effect that we find decay over time? Or does it grow over time? And again, there have been similar kinds of cases.

And then we'll talk about different concepts of what is a dosage issue? And you can think about it like in the medical trials where the dosage is the amount of some treatment given, and there are some studies that are like that. So anyway, that's just background to get on to the good stuff.

Mark.

DR. DYNARSKI: Okay. Thank you.

I'm going to be talking about the Educational Technology National Evaluation, whose second and final report was released in February, and theirs, by design of this study, began—we began thinking about the design aspects in about 2002, got a design, got it into the field, and then we operated in the field for a couple of years, and then there is a year of follow-up, and so on.

So it ends up becoming—oh, and then there's the IES peer-review process. So it ends up taking some time, but there are two interesting dimensions for the dosage question inside the study. One is that across all the schools and classrooms within the study, inside a specific school year, there are much different levels at which the students were exposed to technology, so we have a within-year kind of dosage issue that's going across the units—but then we did something partly at the urging of the software companies who were participating in the project.

We asked teachers to stay in the study for one more year as treatments or controls. They then received another cohort of students, as I'll explain. They had to stay in the same grade level in school, which turns out to be a fairly small number of teachers actually for whom that's true.

If they did, what happened was the study shrunk in size, but now you had an experienced teacher who had been through the whole year using technology, and the hypothesis is that with a year of using the technology, they better understood how it fit with their curriculum; which modules they want to use; how much time they want to spend.

So those two components of dosage, then, form a kind of the content of this session, which is students get more, and then the teachers are more facile, we hypothesized, and able to generate large effects.

But the basic message, in case there's a fire alarm or something like that, is that the results are entirely mixed. You cannot actually see the dosage. You see it; you don't see. You see it; you don't see it. It's just not a very clear thing. I think this generally characterizes a lot of if one looked across a lot of these larger scale evaluations at dosage issues: it never emerges as a clear thing.

What does emerge as a clear thing is that everyone thinks dosage matters. Time and time again, in any kind of discussion with educators or policymakers, they think you just need to do more of the thing, but the evidence is not supporting that statement that I know of.

So here's a synopsis of the study design. It's got a lot of different facets like a lot of these large-scale studies, but essentially there are 15 different products being implemented, nine of which were in reading for younger kids, and six that were for math and for older kids. So we had first and fourth grades for reading, and then sixth grades in algebra for math.

And they essentially operate as four separate studies because even though it’s the same organization doing the research, they’re different school districts; it’s different products, different elements involved.

The four studies are essentially reported out in chapters, and if you took any one of those studies, those study chapters, out, and you didn’t look at the other three, you’d still have a whole full-fledged study there.

There were 132 volunteering schools, and, in fact, volunteering is an important theme throughout here. The districts volunteered, schools volunteered, teachers volunteered, and then there was random assignment.

The kids didn’t volunteer, but, you know, they just get assigned into the classrooms then, but the assignment of kids into classrooms occurs after—rather occurs before the teachers know their assignment. There’s no possibility that the treatment assignment affected where the kids were.

Each school is an experiment, however, because in each of the schools participating in the study, we randomly assigned schools or teachers within that school into the treatment or the control groups.

We had to worry about whether or not there might be contamination issues should the use of the product lead to changes in teaching practice; and later, the Institute began a study to try to really get a handle on whether that contamination issue is a serious empirical one.

But partly because it was informed by the idea that teachers actually don’t talk with each other that much. I mean certainly it’s theoretically possibly that there is contamination, but it does require that teachers are sharing quite a bit and then changing what they're doing; and in the case of technology, it's very easily contained because often there are log-in procedures and controls on who is using it, you know, signing up for the lab, and so on.

It's not so easy for the other teacher to do the thing that would make his or her instruction more effective and thereby reduce the treatment effect, so we didn't worry that much about that.

The companies then trained the teachers during the summer prior to start-up using conventional approaches that they used for getting the teachers familiar with the technology and capable of using it; and at the end of that training session, we gave little questionnaires to teachers asking how well-prepared they felt; and about generally in the high 80 to 90 percent of teachers felt the training got them ready.

Interestingly, by the first time we did a classroom observation, which would have been in about October of that same year, only about 60 percent of teachers felt like that same training had prepared them.

So what happened was, once the rubber hit the road, it wasn't clear at all that the training had gotten a bunch of teachers ready.

We did purchase some upgrades because what we didn't want the study to end up becoming was a place, where you tried to implement the technology and half the computers weren't going to function correctly.

So that might actually depict the state of real schools out there, but it would be a poor representation we thought of a national study that

Congress had mandated under No Child Left Behind. So we felt like ‘let’s make sure that there are some minimum standards being met here by the software.’ To that respect, the study has some efficacy elements mixed inside: what’s overall an effect in this study?

In the first year, the key findings—because I want to talk about the dosage issue more than the key findings, but we didn’t—we tested kids in the beginning of the year and in the end, and test scores in the main are not statistically different between classrooms where the software is being used with instruction and the classrooms where it’s not. It actually doesn’t matter ultimately which product we were looking at.

And then we also did an analysis of whether contextual factors caused different treatment effects, and we’ll look more at that later. But essentially you can, within a hierarchical model, you can imagine the treatment effect being modeled as its own function of lots of characteristics of the teachers and schools; and those characteristics just did not predict the treatment effect.

And then this last result, which is experience. Teacher’s experience proved to have mixed effects on effectiveness. So when the teachers are a year learned in the program, so to speak, that does not mean that they are better or that the kids learn more.

Let’s do the student part first. Okay. So here’s the first year. There’s four sub-studies here, but there is really no reason to repeat the analysis, so let’s choose first grade. These are reading interventions, teaching kids vocabulary and phonic principles, phonemic awareness, and the like.

And this diagram is showing you—we asked the teachers about how many hours they were using technology with their instruction over the course of the school year, and this is in treatment and control classrooms. And now we're going to difference it.

So what you're looking at is the net dosage of technology, and the reason we have to ask the control teachers this is because computers are virtually ubiquitous, and consequently you have to be careful not to imagine that the control kids are getting zero technology.

That's virtually no school in America has no technology in the classrooms anymore, as NCES's surveys have indicated.

We do know—this is a quantity differential. We also know from our own discussions and interviews with teachers that there's also a quality differential because in the control classrooms, the computers are more often used for browsing or routine kinds of word processing functions, whereas, in the treatment classrooms, they're using it more for instructional purposes.

So, in some sense, this under-represents what we think of as the dose, but it's much more easily quantified.

And what we have here—again, it's a very large study—you have 14 different school districts participating with reading instruction. Each one of the numbers you see there is actually a school district school. So, for example, looking all the way to the right, there are four or five—I can't actually tell because the numbers run together—it's about five different schools in district 14.

And so what you’re seeing is the net number of teacher-reported hours of technology use, and variability is a big aspect of this picture. For example, district nine, which is that light blue kind of in the middle, occupies a pretty high—is a pretty big dosage. Other districts like district 11 is not much. It’s about 25 hours. So it’s going both up and down across districts. It’s also going up and down within districts, and so you get a fair degree of had this been a flat line, looking at dosage relating to treatment, it wouldn’t have worked out because there’s no identification.

So with that as a backdrop, now let’s map on top. Now, let’s take a look at effect sizes. These are now the measured differences in learning gains between the fall and the spring. There’s a couple of noticeable features here. One is that it clearly visually centers on zero.

But it’s also quite interesting that there are huge within-district school swings. So, for example, that same district 14—there are six schools there—had the school district done its own evaluation, using these schools as an experiment, it would have noticed that its schools went from a .5 effect size to a minus .5 effect size, which is just a stupendously large difference across schools in what is actually the same product being implemented in different places.

So sometimes it worked, so to speak, very, very well, and other times, it went way in the wrong direction. Now, one can speculate on the reasons for that. One thing we’re going to do analytically is we’re going to try to take the characteristics of the districts and the schools, and we’re going to use regression methods, and we’re going to say if you go back one, I have all those variabilities in things like dosage, and I have variabilities in effect sizes.

So I'm going to look for a quantitative relationship between these. Just to keep a notion in mind, though, you have small numbers of teachers in each school, generally, not more than five at a grade level— five or six. In fact, that would be a pretty big elementary school that had six teachers at a grade level.

Another thing to consider here is: suppose your teachers simply had much different value-added or different ability to produce growth, and then you randomized them? Well, even with no treatment, you'll get large differences in effect sizes across your school, but it will be completely just the teachers either moving into the treatment group or into the control group randomly.

So it's something to keep in mind. That would actually mean that your regression would have zero predictive power because it's entirely just the characteristics of the teachers that's causing the effect size.

We did the three-level modeling, and with treatment being predicted by all the factors we could identify in the school. We had a lot of factors. We had: is there a technology coordinator? Did the teachers receive professional development on how to use technology? What are the ethnic and racial characteristics of the students? And so on. Those are all predicting the treatment effect.

Three out of four times, dosage, just a pure minute difference that I just showed you, has no statistical relationship to that measured effect.

So, in one grade level, it does, which goes back to my basic message, that these are mixed results. I'm not sure I would look at that and say, okay, I see the relationship. It happened once. Even when it did happen, it's a nonexperimental result. So one has to be a little careful about how far one would push that.

So we leave this one saying not proven that there's a dosage relation.

Okay. Now let me just set the stage for a minute here. So now you take that study I just did and now all the kids, it's the end of the school year. All those kids flow out. So they flow to second grade. New set of first-graders come in. If the teacher stayed in the same school, in the first grade, I keep the study going. I do the same assessment pre and post. By the way, those effect sizes were based on the SAT-10 reading test.

I'm going to do it again. But now I have the same teachers in the second year as I had in the first, so I can go back; and the sample will shrink a lot because every teacher that moves on, retires, gets assigned to a different grade level, finds a job in another school district; they all leave the study. I'm only left with survivor teachers.

But I can go back with those set of surviving teachers—that doesn't seem like quite the right term—but, you know, researchers talk like this and we don't worry too much about it. But I have these teachers who are still in the study, and I can go back now and say, well, what effect did those specific teachers have in that first year? Let me compare it to the effect that I see with a fresh set of first-graders in the second year.

And here's what I found in reading. It went down. Okay. It's not statistically significant, but the second year effect that I measure of using technology is, as you see, is actually numerically smaller than it was in the first year.

In the fourth grade, which is also a reading study, but reading comprehension kinds of programs, it goes the other way. But I lose a lot of teachers in the fourth grade just for happenstance reasons. I can't say that is statistically significant, but again, this is a very mixed point of view.

So I not only have experienced teachers, I have their effects in the first year and their effects in the second year. The hypothesis that a year of experience matters I would say is not proven from these data.

In fact, it's a bit challenging to even explain how you could get a negative change since, if the teachers just completely stop doing it, they would become control teachers, and you'd have a zero. So how you would go from a positive to a negative is a little bit—it's not so easy to actually construct a theoretical story that the teachers are worse somehow in the second year than simply not doing technology, in which case it has to be zero.

Here is what happened in math. You'll notice here's reading; here's math. The picture is kind of striking and the exact same thing happens. It goes down in sixth grade. So these are pre-algebraic skills that are being tested, and in fact, in this case, it actually goes down by a statistically significant amount.

And in algebra, it actually went up by a statistically significant amount. Well, statistical significance aside, it's a mixed test of the hypothesis, so not proving this notion that a year of experience will matter.

Some of what I mentioned, for example, the fact that you can only do this with a group of teachers who stay in the same classrooms for that grade level, these are the kinds of considerations that we wanted talked about afterwards. These are not parameters that are very easily known at the outset of a study.

These are things you must power to, however, if you want your second-year study to have enough statistical precision. You're powering to something which is very difficult thing to know, which is, how many teachers are going to be in that same grade level? As a practical consideration, this is a very nontrivial one.

Let me end with a couple of thoughts. One is planned dosage variation studies are feasible in a conceptual sense. You can imagine having assigned a teacher to do a very heavy dosage of technology. And in that same school, this teacher might do light or zero and start building in this kind of theoretical way to test it with a lot more of the design control of a study behind you, rather than just letting it all play out in the data as in that first piece I showed you.

The problem is that in real implementation of that kind of study, it will turn out that the schools have a lot of practical limits on what they're going to be able to do for you in terms of getting the dosage up.

So, for example, in a reading classroom, if you want the teacher to do it a lot with kids, you're going to have to put in more computer terminals so that more of the kids can cycle through the instruction, but where are you going to put them? And in this case, these are first graders. You're certainly not going to send them to the computer lab; you'll never see them again.

So these very school-oriented considerations start to enter here. You can't set up ten computers in the back of a first-grade classroom. They are doing other things with it.

So as much as these kinds of planned dosage studies are interesting and compelling to researchers, in the practical setting of a school, they really start to hit against just pure physical kinds of constraints, and at the higher levels when you get to middle and high school, nearly all of these products are going to want to operate through labs. Lab space turns out to be a crucial constraint on your ability to do all this stuff.

The second point, at which I want to end, is basically the efficacy trials, not necessarily trials which are being done by the Institute, but trials which are being done by the developers, really seem like an appropriate setting at which one could be trying these. Take a few schools within a district; for example, try varying these things, not withstanding what I just mentioned about the practical constraints of working within schools.

I think the kinds of dosage studies that we hear of from health research, those are all being studied prior to the time when you get to the very large-scale effectiveness study like the kind of study that we did here.

What was interesting, however, is that when we talked with the companies early on, they really had very little evidence backing their models. We would ask them, you want us to implement your program? Tell us about your model.

And they would say, “well, it’s three times a week for 20 minutes a day or so on,” and we would say, “so what kind of empirical support?” What if they only did it 2 days a week? And they’d say, “oh, that’s fine, too,” and what if they only did it for 10 minutes a day? That’s really up to them.

I mean—

[Laughter.]

DR. DYNARSKI: And so in the end, we realized they didn’t have a model. They had a plan. They didn’t have a tested advocacy-driven model that said, okay, if they don’t do this, you won’t see effects, but if they do that, you’ll get effects. It’s just not that clean.

And so, consequently, when we went to implement and then we began hearing noises like, gee, we’re worried about fidelity, we thought fidelity to what? I mean there is no evidence that drove the original model of implementation. What would you have fidelity to?

So, in the end, we really thought, you know, we need to keep—there’s a very nice process within drug and health services research by which ideas go from small pilot, you know, let’s test it to see whether something might happen, and it builds upwards towards these large-scale effectiveness trials.

The effectiveness trial is not the place at which one should be figuring out what one should have done in the first place. That's what we see. So let me stop there, Rob.

Thank you.

DR. HOLLISTER: Thanks.

[Applause.]

DR. HOLLISTER: I like that remark about if you send a child to the laboratory, you'll never see him again—computer lab. So it's a new theme on the old one, the dog ate my homework; the computer ate my child.

DR. DYNARSKI: Yeah, yeah.

[Laughter.]

DR. HOLLISTER: Affected the attendance. Okay.

DR. DYNARSKI: Patrick, let me get you going. Okay. You're up.

DR. WOLF: Thank you, sir. Well, thank you, Rob.

I appreciate IES inviting me to present here. When I first was extended the invitation to make this presentation on a dosage in the DC Choice, I figured, well, heck, I'll just do a dramatic reading from several appendices of the report.

[Laughter.]

DR. WOLF: But then the editors of *USA Today* described this report as impenetrable.

[Laughter.]

DR. WOLF: And on the off chance that one of them is here today, I decided instead to do a PowerPoint presentation.

This is a presentation of results from a collaborative study sponsored by IES, supervised by IES, and thanks to Phoebe Cottingham and to Marsha Silverberg, who have been tremendously supportive in this effort.

It’s a collaboration between Westat as the prime contractor, me, and my team at the University of Arkansas, Chesapeake Research Associates. I see there’s great turnout from Chesapeake Research Associates today—

[Laughter.]

DR. WOLF: —in the person of Mike Puma, and one of my former colleagues who’s, one of my former Georgetown University colleagues, Nada Eissa. There also are, there is representation in this room from our elite Technical Working Group. I thank them. I will not call them out because they’re currently in the Witness Protection Program.

DR. WOLF: Okay. So today I’m going to talk briefly about the program; about the study; about the achievement impacts we observed in year 3 that were released about 6 weeks ago; about achievement impacts over time, because that’s essentially the sense in which a school voucher program has a dosage component; and then I’m going to talk briefly, as Mark did, about dosage in the context of this particular educational intervention.

The Opportunity Scholarship Program was established by the DC School Choice Incentive Act, which was signed in early 2004. It provides $14 million a year for approximately 1,700 students to receive vouchers worth up to $7,500. And they can use these vouchers then to attend a participating private school of choice.

To be eligible for the programs, students must be DC residents; they must be entering grades K through 12; and they must have family income at or below 185 percent of the poverty level; essentially qualify for the Federal Lunch Program.

The statute included a prioritization scheme whereby applicants from schools designated as in need of improvement under No Child Left Behind were required to be given some priority in the award of scholarships.

And through the 5 years that the program has operated, a total of 68 private schools in the District of Columbia have participated.

The impact evaluation is a randomized control trial that draws upon the fact that baseline-eligible applicants were assigned to receive a scholarship offer or be assigned to the control group based on a lottery. This was for the years and the grade spans in which the program was oversubscribed, which were grades six through 12 the first year, and all grades, K through 12, the second year.

So our impact sample is comprised of these first two cohorts of participants in the program: Cohort 1, who applied and were randomized in the spring of 2004; and Cohort 2, who applied and were randomized in 2005.

A total of 1,387 of these students were assigned to the treatment condition; 921 to control. Cohort 1 outcomes throughout our study have been lagged one year and combined with Cohort 2 outcomes for our impact report so that both Cohort 1 and Cohort 2 have experienced the same amount of time since random assignment, even though they entered the program in different years.

Outcomes have been tracked annually, and they include student performance on the SAT-9 in reading and math. That will be my emphasis in this presentation. All the results I present will be the achievement impacts.

But we've also examined the impact of the program on parent satisfaction; views of safety through parent surveys; student satisfaction; views of safety through student surveys; and some of the educational conditions that might have been affected by the program through surveys of private and public school principals.

The analysis includes three estimations of impact. The primary estimation of impact is a regression-adjusted intent-to-treat impact estimate. A simple comparison of the outcomes—average outcome experience by the treatment group minus the average outcome experience by the control group.

We then generate an impact on the treated estimation through Bloom adjustment, basically netting out from the treatment group students who never used their voucher.

And finally we provide an instrumental variable analysis of the effect of private schooling, since we also had some members of the control group who obtained private schooling.

So, for an effect, an estimation of the effect of private schooling, if you think that's the treatment here, which many policymakers seem to, we do provide an IV estimate using the lottery as the instrument.

I'm just going to be presenting ITT impacts here. And here's the summary of our year 3 results. Overall, the program demonstrated a positive statistically significant reading impact of 4.5 scale score points. That's 13 percent of a standard deviation and equates to a little over three months of learning. No impact was observed in math.

We also examined impacts for certain policy-relevant subgroups. We had five subgroup pairs, so a total of ten subgroups, and basically we observed reading impacts at the subgroup level for students who were not attending schools in need of improvement at the time of application; for those who were in the higher two-thirds of the performance distribution at baseline; for those who were female; for those who were entering grades K-8 at baseline; and for Cohort 1.

Now, the female and Cohort 1 subgroup impacts in reading appear in italics because after adjusting for multiple comparisons, those particular impacts were no longer statistically significant.

No impacts in reading were observed for students who applied to the program from SINI schools; those who were in the lower third of the performance distribution at baseline; males; those entering grades 9 through 12; and Cohort 2.

No impacts for any subgroups were observed in math. So really, sort of all of the action on the achievement side from the intervention appears to have been on the reading side.

So here is a graphical depiction of the reading impact 3 years after random assignment. The point estimate is 4.46, about 4.5 scale scores, and you see the confidence interval there is above the origin leading to the determination of a statistically significant effect.

And you can contrast this with the impact estimate for math. Though positive, the 95 percent confidence interval clearly intersects the origin. This is a zero impact finding.

Okay. So that's after 3 years and simple consideration of two datapoints: baseline and 3 years later.

In order to factor dosage into this consideration, I think it's useful to examine the trend over the 3 years of the evaluation. Basically, across the three outcome years we've studied so far, there appears to be an apparent trend of cumulating impacts in reading, but no apparent trend in math.

When we look at the subgroup level, across the three outcome years, of the five subgroups in which we observed statistically significant reading impacts in year three, all but Cohort 1 show a trend suggesting cumulating impacts.

Of the five subgroups without impacts in year three, I deliberately put "x" there. The "x" suggests cumulating impacts because I think it really is a judgment call. Anyone could draw different interpretations about how many of those demonstrate trends, and I'm going to present them for you so you can draw your own conclusions.

The critical thing to keep in mind is that our data are not yet pooled across years, so when we present these results over 3 years, we do not know if the impact in one year is significantly different from the impact in another year. We haven't pooled them and determined the covariance across years so that we could actually do significance tests for the impacts across years.

All of our significance tests to date have been limited to within-year evaluation. So this is going to be an impressionistic presentation of descriptive information.

So think of it as looking at Monet painting, not at a CSI crime photograph. So our first Monet painting is of the overall impacts over 3 years reading and math. And here is the source of my—let's see if I can get this to work. Oh, it's out of range. Probably couldn't even bring down a plane with that.

Here you see on the reading side, the basis of my statement that the trend suggests cumulating impacts over time is that in year one, there was an experimental impact of one scale score point; not statistically significant.

In year two, a little over three scale score points; not statistically significant. And in year 3, 4.5 scale score points; statistically significant. So there are suggestions of a trend of cumulating impacts.

The math side. Math actually looked more promising in the first year, with a treatment group advantage of a little over 2.5 scale score points. Not statistically significant, but there has been no action on the math side since then, and so you might suggest there could be hints of a quick reversion to the mean on the math side, or you could simply interpret that as just zero, zero, and zero.

Now looking at the subgroup level. Only reading, because on the math side it’s really all noise at the subgroup level in terms of trends.

But on the reading side, you see to your left, the trend of impacts for the students who had not attended a school in need of improvement before applying to the program; and you see suggestions of a trend of accumulating impacts over time.

In terms of the students who switched from SINI schools, is that a trend? Year three is higher than the negative in year one and the zero in year two, but there’s not much to draw from there. These may very well be just fluctuations; random fluctuations around the origin.

For the higher performance students, again, we see suggestions of a trend, though perhaps not quite as much of a push in year three compared to year two. In terms of the lower performance students, there were actually negative, but not statistically significant, impacts year one and two, and then a positive, but not statistically significant, impact in year three.

So the directionality flipped, which might suggest a trend, but that would be a pretty narrow read on which to hang any conclusions.

Females compared to males; really certainly a suggestion of cumulating impacts for the female students, and for the males, everything is on the positive side, but there doesn’t seem to be a compelling vision of trend there.

K-8, a trend pretty similar to the overall trend impact for reading; very reflective of the overall average impact across the years; a little more precise in year three. It gets those two stars.

And then 9 through 12. You know, maybe high school is where promising education reforms go to die, but that looks kind of like a dead body there.

And, finally, the cohorts. Cohort 1, the first movers, showed suggestions of positive effects straight out of the box. They got somewhat larger and became statistically significant in year two and three.

Cohort 2, sort of a mini-trend, you might say. None of those impacts statistically significant in their own right, but they seem to be pointing in the right direction.

So what can we say about dosage in light of these results from the DC Choice evaluation?

First of all, they suggest that impact patterns may differ by domain. Basically, we may have seen somewhat of a reversion to the mean on the math side. We certainly saw no evidence of reversion to the mean on the reading side. To the contrary, the reading impacts appear to cumulate over time.

No obvious reason for the difference. We'd certainly entertain suggestions from the audience if anyone has them. It's something we're going to try and look into as we continue this study.

But another point I want to make is that school choice interventions in particular may require long treatment exposure to demonstrate results. And the reason for this is that school choice is an intervention that starts with a school switch.

In order to use a voucher, the students had to switch out of the school they were attending and into a new school with different expectations, different peers, a different environment. All those things could be positive and may have placed them on a positive growth trajectory vis-à-vis the control group students, but it also probably interrupted their learning progress for some adjustment period, and they may very well—we couldn't measure impacts shorter than 1 year out—but within that first year, they actually may have realized a negative impact from the switch and the necessary adjustment.

Other research by Eric Hanushek and others has suggested that's the case, and so particularly when evaluating school choice interventions, drawing policy conclusions from them, it does appear to be important to give the treatment a substantial amount of time to mature before drawing strong conclusions.

We have authorization funding for fourth year analysis. We're in the field right now wrapping up data collection, and so we'll have a fourth datapoint on the trend line in about a year, and like bring it on.

All right. Thank you.

[Applause.]

DR. WOLF: Oh, the last thing, one thing I forgot to say is that I was only speaking for myself. So none of my opinions, interpretations or jokes should be attributed to IES or University of Arkansas or my patient colleagues or anything else.

MR. CORRIN: Hello, everybody. So I get to wrap up, I guess, the presentations on this distinguished panel, and I just want to say, Patrick, that your report was actually read by the popular press and that counts for something. I don’t think the report that I’m reporting on was read by not too many people, but we’re not there yet. Debbie Viadero did a nice little blurb for us.

So I’m going to talk about a project I’ve been working on for about 4 to 4 and a half years, the Enhanced Reading Opportunities Project, and it’s been a great project. Paul Strasberg and Marsha Silverberg are here from IES, and they’ve been with us the whole way. I’ve got TWG members in the room that have helped us the whole way and various other people who I know have actually given me informal and formal feedback. Thanks to everyone.

Okay. So briefly, I’ll give you background on the project itself, and then the question I currently focus on isn’t so much on dosage for students, but rather for teachers. The teachers need more than a year to master new practices.

So I’m going to talk some about our implementation findings; a little bit about our impact findings; and then I’m going to look at two kinds of comparisons: teachers who taught this for 2 years versus those who came in the second year and replaced earlier teachers; and then how did those two-year teachers do first year compared to second year?

Okay. So this study, it’s an impact evaluation of two supplemental literacy programs targeted at ninth grade students, and the main research question we’re trying to answer is an impact question. So really we wanted to know particularly the first 2 years of this study, if there was an impact on reading comprehension for these kids; and also we looked at some reading behaviors through the student survey.

The two programs are something out of WestEd, Reading Apprenticeship Academic Literacy, which is sort of a pullout class version of their schoolwide model Reading Apprenticeship, and then Xtreme Reading from the University of Kansas Center for Research on Learning. It's based on their strategic instruction model that they've been doing for many, many years actually and continue to refine.

These classes are intended to replace electives, and they are. When we went through the selection of these programs, the panel debated how much they want things that are two programs that are very different, as opposed to two programs that are similar. I think what we ended up with is two programs that share a lot of common philosophies and principles, but I would say the implementation style is different.

The Xtreme Reading Program is a little bit more structured. And the Reading Apprenticeship Program I think offers a little more flexibility to the teachers.

But the intent was for these to represent kind of class of intervention: So what happens when you give a supplemental literacy class to kids who come in behind in reading?

There are two cohorts of kids. The first cohort of ninth graders was '05-06, the second '06-07. There is one ERO teacher per high school. There are 34 high schools that participated across ten school districts. They were respondents to a grant application that actually came out of OVAE at the time and now is managed by OESE.

And they were not targeted at reading teachers. These are typically English language arts teachers and history teachers who are trained specifically to deliver this reading class.

And then section size, relatively small, and then, I would say, fairly substantial support was provided to the teachers along the way in terms of summer institutes; school year coaching and onsite visits; off-site booster training; and then again in the second summer, you had a ramp-up training or another booster and then further support.

So during these 2 years, starting the summer before the first year of implementation, implementation through the 2 years, there's a fair amount of professional development.

So we collected some implementation data, and for that I have to give credit to AIR, in particular Terry Salinger, who sort of ran that effort and sort of oversaw a lot of that data collection. So I get to stand up here and talk about it, but her team did a lot of the work with some support from MDRC.

But we were curious how well these things were put into place. We were curious about how this data also allows us to look at one of the things that came up initially, this project was set up to be a one-year project. It was actually expanded to be two cohorts of teachers. So we want to see if implementation changes over 2 years? There's an expectation, which I think Mark referred to, that the more you get it, the better you get at it.

And then we also did some investigations, nonexperimental investigations, of relationships between implementation and impacts.

And then implementation fidelity. There’s basically a site visit a semester starting the second semester of the first year, and we looked at two things that the program developers cared a lot about in their programs.

One is what kind of learning environment do you set up in the classroom? So is it a place where kids really should be able to learn, and that you’ve got good relationships there instructionally, but also are you delivering instruction that’s going to deal with reading comprehension?

And then there’s a variety of constructs that build into these measures, and in the end you come out with this number of one, two, or three essentially for each of those two dimensions.

All right. And then we did one other thing, and this comes into place now as I start talking about implementation. We categorize schools based on these averages. We called places that were well aligned, having a two or higher on this three-point scale. Moderately aligned, 1.5 to 1.9, and poorly aligned, less than 1.5.

In particular, we’ve made greater distinctions at the lower end of the scale because we were really trying to see where are the places that really struggled with this and really suffered through this, as opposed to places that seemed to reach sort of a moderate or well-aligned threshold.

Okay. So here’s what we learned. In the second year, which is what the report was about, we had 34 teachers: 25 that taught the entire first year of the study; two had taught a portion of the first year, which is actually most of it. Usually these two teachers were replaced into the second semester. And then seven were brand new at the beginning of the second year.

All the teachers that started at the beginning of year 2 finished year 2, which is great. There may be one exception of someone who had a 6-week maternity leave, but basically everybody taught the whole year.

What we saw in the second year is when we looked at classroom learning environment, when we were focused on where the places that really struggled, poorly-aligned programs—in the second year, there was only one school that fell in that category versus four in the first year.

When we looked at reading comprehension instruction, we only had one school fall in that category compared to nine in the first year.

So it seems like from a categorical perspective, implementation was better in the second year. When we combined these things and look at them overall, we also had 23 schools as opposed to 16 in the second year that did stuff well. So we didn't just look at that bottom end.

And we saw similar findings across both programs, so the fact that there are some differences in how they were implemented didn't seem to play out in terms of whether teachers did better or worse with this or schools did better or worse with this in the second year.

When we look at averages and we look at a continuous measure, again, we see that things are higher in the second year as opposed to the first. I mean the general sense and what we got from the field is that things went better in the second year. The programs were implemented better in the second year. So sort of this assumption of if you have more time to do it, well, you do it better seemed to play out.

I'm going to—here's bar charts across the three site visits. Learning environment, you've got overall, and then you have the two programs. You know it's a little bit up and down. There's less movement there. But when we look at reading comprehension, that's a place where we actually see more of a linear kind of pattern of results.

Actually the end of our first report, we did a comparison, and we looked at two subgroups of schools. We looked at school that ran the programs for longer during the year. There's some places that started up later and had higher fidelity versus those places that didn't. And we just looked at these two groups, and what we saw with these two groups is that the schools with better first year implementation had higher—statistically higher—impacts than the schools that didn't.

So we may even have a little hint at the end of the first report that says, you know, things have started in the second year, and it looks like places started sooner, and teachers are doing better. So our prediction was, hey, if this continues, maybe we see better impacts.

All right. So I'm going to hold off and tease you for a second. Random assignment design. 34 schools randomly assigned; ten districts; 17 to each program at the school level, but then within each school, there's student-level random assignment, to whether you stay in an elective class or whether you're in the literacy class.

Okay. And this is what we see. Reading achievement measured with the GRADE reading assessment. Both years we found overall positive impacts statistically significant at the .05 level of basically the same size.

So needless to say, with our second cohort, we were mildly disappointed, because we felt good about how the programs were going in the second year. In neither case did we find impacts on vocabulary. All right.

So what we see is, yeah, implementation got better and we think the schools actually learned how to do the programs better, but it didn’t translate into a different impact with the second cohort of kids.

Okay. In terms of those reading behavior outcomes, we didn’t have anything statistically significant with the first cohort. We did in the second cohort have one on reading strategies. We asked the kids to report on how often they used the types of strategies that are taught in these programs in their content area classes.

However, when we look at an omnibus test, when we grouped the domain together of reading behaviors, that overall test isn’t statistically significant, so we interpret this with caution.

But I’m still going to talk about it because this is my hypothesis-generating presentation. We did some exploratory analyses because we found all this stuff pretty interesting. We looked at whether implementation changes over time? Like I said, we were curious what happens over the 2 years, but also how do these teachers who return compare to the teachers that were a replacement?

One of the things that we see consistently is that from the first year, the second year, typically we got higher results. The one thing that you can see: that year one the schools with replacement teachers—these are the teachers that left essentially—in that upper right-hand cell—tend to have the lowest scores, and it's the same for the reading comprehension instruction. And when you look at year 2, the scores are pretty much the same between the replacement teachers and the teachers that taught the program for 2 years.

So, sure, you know, it looks like implementation gets better in the places where teachers had more time with it, but also you have these replacement teachers come in. Their first year they do pretty well. I'll talk more about that in a minute.

Overall implementation fidelity, again it's just more evidence that in both these types of sites, those with 2-year teachers and those with replacement teachers—here's a little graph—you see everybody to the left side of that diagonal, or the vast majority, everybody did better in the second year.

So then we say okay, what about impacts and how those look for those different groups? Again, we see right here in the second year, if you just compare the replacement teachers and the 2-year teachers, and I should really say the schools with these two groups, the differences are not statistically—the impacts are not statistically different between these two groups even though they are significant for the 2-year teachers—okay—in the second year.

And then when we look at just those places we have replacement teachers, again, in year two, we see relatively similar impacts. The reading strategy is one—that's the one place where there's a difference that's statistically significant.

So I'll summarize and then—I'll summarize and say a couple other comments. Basically, yeah, we learned this could have a positive impact. We were very excited about that, but it's a pretty small impact. The programs, what we saw, particularly by the end of the second year, were implemented with reasonable fidelity. They looked kind of like what the developers would want them to look like on average.

Generally it was better in the second year. Now, when we talk about the second year, I've talked a lot about teachers, but really there's one teacher per school and the school is implicated in this. So I think this isn't just about whether the teacher is learning about doing the program better. There are other staff in the schools that are learning about "how does this fit into my school?"

So an example of how that might work: in the first year in some schools there is sort of a basic set of supplies you're expected to purchase with your grant money. In some schools those supplies are ready on day one. In other schools, they trickled in over time, and that's about a logistic within the school, getting these supplies to the teacher's classroom.

Then I would say, again, you know, it looks like those teachers that returned did better the next year, but at the same time, you've got these replacement teachers come in, and they do as good a job in the second year.

So again, we can hypothesize that some of it has to do with the school learning to do things better, but also the selection of teachers.

So I think one thing we know about the replacement teachers compared to the starting crew of teachers, they were more rigorously evaluated when those resumes came in for who was going to teach this program.

I think also it speaks to the school in terms of what the school understands about, hey, we had this teacher that really didn't work out so great for us; I need a different kind of person to do this.

So there's a combination of both sort of on the associations in terms of OVAE reviewing some of the or OESE reviewing some of those resumes and drawing a little bit of a stricter line, but also with the school I think learning. Something didn't feel right.

And then, you know, lastly, we have sort of the impacts. The impacts stay about the same. So what's the association between what the teachers are doing and impacts is sort of an open question.

We have this one thing on reading strategies. I leave it in there. Again, it's really hypothetical, and it's sort of cautionary, but if there's one thing, I feel like if I'm teaching the program better. All right, the thing that's newest to me as a teacher. So our teachers averaged about 11 to 12 years of experience when they came into the program.

I've had to set up a classroom. I've had to work with kids before. I've had to set up relationships. So if there's one thing I've got a little less learning to do, it's on this learning environment piece.

I may be tweaking that to be a little closer to what the program wants, but this reading comprehension stuff is new to me. So here's a place where learning more about the program may really affect—having more experience with the program may affect this.

So, you know, seeing something for reading strategies for these returning teachers, it kind of makes sense. If this was new to them, that maybe the kids—they've learned better how to get those strategies to the kids.

But I will stop at that because we want to have some time for discussion. Thank you.

[Applause.]

DR. HOLLISTER: Okay. We'll take some time now for questions on any of the presentations. Go up to the microphones if you have some question that you want to pose and tell us who you are.

DR. YARNALL: Hi. My name is Louise Yarnall. I'm from SRI International.

I wanted to ask Patrick Wolf, in particular, I was struck by the talks at lunch and breakfast this morning about the need to have data and findings that are interpretable by decisionmakers and policymakers. Based on your data, what would your professional judgment be; what should you do with the school voucher program?

[Laughter.]

DR. WOLF: With all due respect, I don't really see that as my role. My role and the role of the team is to bring the data to the conversation in as reliable and clear a way as possible, but certainly I'm not going to recommend any specific policy action based on these findings except, you know, keep funding our study. Maybe I'll recommend that.

But to your point about results being understandable and interpretable, I think critical to us communicating our results in our third-year impact study was the conversion of the effects to approximate additional months of schooling. I think that's been very helpful for people to get a sense of the magnitude of these impacts and what they mean in a sort of real educational sense.

DR. HOLLISTER: Yes.

DR. SUPPES: Patrick Suppes, Stanford University. I thought these were excellent presentations, very clear, and I think I have a reasonable understanding of what happened, and I like that.

My question is about the methodology of dosage. Now, of course, in statistics, dosage is an old subject. It has a huge literature, quantitative in character, but I also think that there are other models, and I want to suggest a different model, but I'll say what I'm talking about first.

And thing that is, let me introduce the word "work" or the concept of work. I think one of the disappointing things in the report on the technology here, that even though there's a lot of computer information that's been collected, there's no report of actually what did the students do?

For example, I mean in recent studies of EROs, we found that much more important than whether they're in the experiment or control group is how much work did they do? For example, I mean one of the measures we've used, where we can measure it on the computer, is correct first attempts at exercises, the actual number that the students did.

I also want to emphasize that in education we overemphasize time. As any physicist will say, time is not a causal variable. There's got to be some activity, and you should measure that activity, and it seemed to me with these very detailed and beautifully presented studies, it's a shame that there isn't an additional measurement, attempt to measure what do the students actually do, where you actually have the possibility of looking at a great variety of exercises.

My own experience is, which is over very many, many years, that regardless of whether a student is in the experimental or control group, there will be huge variation in students in this, and our own data with fairly good numbers show that very much.

So it seems to me that there's a real opportunity, instead of using the word/concept of dosage, which is fine—I don't object to that per se—but I much prefer a physics model to a medical model; it's less primitive in its conceptual apparatus—is to have a measure of work and to attempt to report on the differential because what we find is that in experimental cases where the students are working a lot—that may happen in the control, of course, you've got the same thing to watch—you may have very big effects, and we have some very big effects from that measure.

So I'd be interested in comments on that piece on that.

DR. DYNARSKI: Well, let me comment that in general because this is a national evaluation that's taking place on a group of products in many districts and schools, we have to use concepts that will translate across different kinds of settings like that, and especially with different kinds of reading standards and reading curricula across states.

The notion of measuring, say, work in reading is actually one that would require, I think, effort beyond what we could have invested here because of the nature of what it means to measure it—is a very micro kind of assessment.

So what we did was we opted for a strategy in which classroom observers could go in and with not very much inference try to assess the extent to which technology changed what the teacher did, which is not the same as the measure of work that you're suggesting, but it's basically empirical evidence that technology did not leave the classroom unaffected.

So, for example, we have measures. We have time sampled measures of the extent to which the teacher was working with groups of kids as a facilitator, versus lecturing. And so the hypothesis in the technology community is that lecturing is not as effective as small group interaction, and so—and we saw very large effects, if you want to call it that, on the geography of the classroom in that respect.

And then likewise, we also are assessing using the same time sampling method; whether the students are on task according to the observer.

And these are trained observers with reliability, and the students were basically about the same degree of being on task except in one of the four grade levels. So I agree with your point in general. I just think that what's being called for is probably something which is better done within that kind of efficacy setting where one can take a very close look, nearly daily, at the classroom activities, but the constraints of these very large-scale studies make it very daunting to try to do that in very many settings.

DR. WOLF: I think that your question certainly speaks to an intervention like school choice, too, because the actual manifestation of the school choice intervention is going to be very heterogeneous. I mean that's the whole point, is to give parents choices, and they'll make different choices and choose different schools, and how each treatment student experiences the treatment is going to be very different.

But at the same time, and we do explore that a bit in our study. We do an exploratory analysis of the effect of the scholarship offer on various school conditions for kids, and there are some sort of tentative findings.

I mean the strongest one is that the students offered vouchers attended much, much smaller schools, schools that were about half the size of the students in the control group.

But other than that, there isn't much that's really strong or convincing. I would just defend, you know, the application of the randomized control trial in this case because I think with the school voucher evaluation, where you've got this intervention that's highly controversial and where self-selection bias is presumed to be a great threat, I mean I think you really do need to avail yourself of the great, you know, leverage and the great protection against selection bias that randomizing on the scholarship offer gives you.

And if that means you don’t learn a lot in detail about, you know, contingencies down the road, you know, that’s certainly something I’m willing to accept.

MR. CORRIN: Yeah. I mean I guess I would just add, you know, I think conceptually it makes, what you’re suggesting makes a lot of sense, and the degree to which it can just be considered for whether it’s an efficacy trial or other kinds of studies to know what the kids are doing is important.

I mean one of the notes I wrote to myself as I was sort of thinking about these presentations is, you know, you have quantity and you have quality, and you sort of want to know an interaction of it, right? So are you giving kids a whole lot of time where they’re not doing anything or not really getting anything that makes a difference for them versus a whole lot of time where it seems like they’re very involved and they’re doing things that seem to push them forward?

I think there’s a lot of good measurement questions about how you capture that and how you do it within resources of a study, and so on and so forth. But conceptually, I think it makes sense that it’s not just quantity.

DR. RABINER: Hi. My name is David Rabiner from Duke.

I have a question for Mark; actually two questions. One is, I was curious as to what the companies whose products were tested, what kind of response they have had to the essentially negative findings, and what kind of criticisms or concerns they have about how their products were tested?

And then the second question is, one would certainly hope that if the results were more positive, that it would have led, it would have contributed to greater adoption of these technologies in schools.

Given that the findings were not particularly positive, is there any indication that schools that were previously using these approaches are discarding them for other things?

DR. DYNARSKI: Let me try to answer your second question first. It's the easy one. I have no idea whether, what the schools or the school districts did. I can say that going into it, they were very eager to participate because we actually approached districts who had indicated an interest in purchasing the products to the companies but hadn't yet done so.

And so the companies relayed the names of those districts to us. We came in and said, hey, how about 2 years for free, only half the teachers, but all the rest of it is a package, and so that's a pretty winning combination for a school district because school districts pilot lots of things, and so we were essentially saying let's do a pilot, and we'll pick up all the costs except for the fact that it has to be for the teachers and only half of them, and they were—we sometimes hear that random assignment raises ethical issues.

But the word "ethics" never actually came up in that setting. They just didn't, it was sort of all plus to them.

With respect to how the companies reacted, individual companies I don't know of any reaction. I do know that there was several statements from technology constituent organizations like the ISTE and CoSN—I'm not sure—I forget what they stand for—who essentially adopted the strategy that we knew this already, which is a time-honored one when you do a national study at a scale no one has ever attempted before, and they say, “we already knew that, and you think, how could you have known? Nobody has ever done this before.

But they essentially said that implementation, we always knew that how schools implemented these things would really matter, that you needed school leadership, you need dynamic principals, you need continual supportive teachers, and the like, and this, this led to an interesting exchange at an American Educational Research Association conference presentation in which my discussant on the panel that day said if they actually had known that was as well-held an empirical regularity as they said it was, then they shouldn’t have been selling the products to places that didn’t have those characteristics.

I never heard a response to that. That was not what we said. That was the discussant who basically said that if you know that it doesn’t work in a place with poor leadership, then ask before you sell it. You know, how’s the leadership here? But they sell it to anybody who buys it.

DR. HOLLISTER: David.

MR. JUDKINS: I’ve got a question for you Mark and also one for William. The one for you, Mark, is that my experience is that given the sort of dosage information that you got from the teachers here about hours is sort of like pulling teeth.

I mean how badly did you have to torture them, them and the data after you got it from them? How much editing and imputation?

And before I step away, the question for William is .09 seems like an awfully small effect size to have detected with your sample size, and I’m wondering what magic secret you had to extract that much information?

[Laughter.]

DR. DYNARSKI: So do we. With respect to how we got that information, that’s actually—David, that was we asked the teachers basically for within a reference period, which was basically in the two weeks prior to the interview, how many minutes did they, had they actually slotted for the use of technology in their classrooms?

What we also gained from the software packages themselves was that software packages count logged-in minutes. You can actually get a validating measure of usage from that. These numbers will not be the same because the logged-in minutes sometimes are being shared by students, and also if the teacher sets aside say a class period for using the software, to the extent that students are absent that day, they don’t have logged in minutes so the number always looks lower.

In some sense, logged-in minutes is a very accurate representation versus what the teacher said that they set aside which is often more of a planning parameter.

However, because we were essentially looking at net dosage across the same question being asked in treatment and control classrooms, we were a little bit relying on the design to get rid of some of the arbitrariness of the measure.

MR. CORRIN: I guess in response to your question, the SAS on my computer has a function called “proc fabricate.”

[Laughter.]

MR. CORRIN: No, that’s the flip answer. The real answer and the part of the presentation I kind of whipped through is, in the design, impact estimates are based at the student level random assignment occurred. The sample is 2,400 kids or so for that second cohort, and that’s plenty big enough even when it’s blocked within those schools to get that effect size.

MR. DAVISON: Mark Davison, University of Minnesota. And my question is for Patrick actually. Vouchers as such are a distal cause, if a cause at all, and so the question is what do you think are the immediate causes of the, potential causes of the difference? In other words, what are the private schools doing differently than the public?

And there may be some dosage issues here. In other words, are they spending more time on reading? Do they have the same length school year? The same length school day? Things like that.

And I have a second question. How did you define private school? I mean there are a number of classes of private schools. There’s the for-profit. There’s not-for-profit, religious not-for-profit, nonreligious, and so forth. So how did you define private school?

DR. WOLF: Sure. I’ll answer the second one first. Basically, we defined private as nonpublic. So all nonpublic—I mean not to be flip. But all nonpublic schools in the District of Columbia were eligible to participate in the program.

In terms of the set of schools that actually did, they were a mix of

Catholic schools had the largest footprint as they do in most urban environments. Other faith-based schools, nonCatholic faith-based schools and some secular schools, including some of the elite prep schools of the District.

So there was quite a bit of variation in terms of the types of private schools that participated. The faith-based and particularly the Catholic schools enrolled the lion's share of the actual voucher students, but there were, you know, nontrivial enrollments in some of the other types of schools.

In terms of—yeah, I mean I could just slightly elaborate on the comment I made about what's happening inside the black box and the real challenges we face in teasing that out and how that's really kind of a secondary concern for us. I agree, it's still an important one, but I'm just not sure we really have the tools and the leverage to tease that out.

What we found in terms of what's different downstream for the treatment and control group, again, the treatment group is attending much smaller schools. They're more likely to attend schools that have programs for advanced students. They are more likely to attend schools that have computer labs and music programs.

On the flip side, they're less likely to attend schools that have individual tutors in the schools, that have programs for English-language learners and students who are struggling in various areas. So they're less likely to attend schools that have certain sort of targeted and defined programmatic supports.

They're more likely to attend schools that sort of have kind of a standard or traditional kind of private school program.

MR. SMITH: Jeff Smith, University of Michigan Economics.

This is going to sound like a series of observations, but you should add a big question mark at the end. And then comment. So I think my appetite was whetted this morning by our thoroughly geeky panel on multiple comparisons, and so I was hoping for more geeky methodological stuff today than we got in this panel.

I mean I liked the presentations, too. I was glad to hear about the DC voucher experiment, and I had heard a little bit about Mark's thing before. But in the course of listening to them, some questions came to mind. The first is what's the question here?

Is the question is there, are there dose response effects? I think we know the answer to that maybe from theory, and so I think maybe the interesting question is when are there dose response effects? What are the covariates that determine whether, what there are dose response effects? And so which margin do they operate on?

So, Mark's margin was an intensive margin. How many hours within a year? The voucher study was an extensive margin. How many years? Right. Well, those are different. Those are different questions—right—different definitions of dosage.

A lot of the standard empirical things that we do have implicit in them strong priors about the dose response effect that it struck me today we don't pay any attention to. So when we run regressions of earnings on years of schooling, which labor economists do constantly, implicit in that—right—

we put in years of schooling. It's this idea that there's a sort of the constant dose response effect or at least in logs.

When we study job training programs, we put in dummy variable for whether or not you got the job training program, and we never pay any attention to how long the job training program was or anything like that. We just put in dummy variable. Well, implicit in that is an idea that, in fact, over whatever range there is of dosages that's offered by the training program, it doesn't matter what the dose is.

Why is that? Are there models here? I think maybe there is theory out there. I don't know this side of the education literature at all. I would like to hear more about theory. And then I would have liked to have sort of jetted off and said, you know, started to think about some of the labor economics theory here. Maybe there is learning by doing versus subjects that are like reading that are sort of the same—you know, the more you read, it's just kind of bigger words, whereas, in math, you know, every year it's completely different concepts. Does that matter for the dose response effect? It seems like there would be interesting things we could talk about there.

And the last thing was statistical treatment rules for dose response effects. So in the years of schooling context, which I think is—we don't think of that usually in the dose response context, but it clearly is, policy is set up where we say you have to, we legally require you to take a certain dose—right—a minimum school leaving age. And then we let you pick what your dose is and we set up prices and we have sort of kinds of things like that.

For some of these other programs, would it make sense to have some sort of statistical rule for trying to assign longer doses to people who benefit more from longer doses or the reverse?

I'll shut up there. Big question mark.

DR. DYNARSKI: What were the questions again, Jeff?

[Laughter.]

MR. SMITH: I think the dose is too big.

DR. DYNARSKI: I can offer thoughts on why do we actually just model these interventions as a zero one in the end. You know you have a highly controlled experiment. But you don't have a highly controlled dosage. You have an endogenous dosage. So to the extent, I mean this especially was true in our study of afterschool programs. It was clearly up to the kids and the families to decide how many days a week they were going to be there.

The only thing the study could really do is say you have access to it. And so in a sense, we have a lot of clear statistical structure at the point of the zero or one and things get very messy at the point of when you introduce the rules or the kind of behavior about why that you chose what you chose.

We did try in the afterschool setting to look across years because we had the same, young kids in the study for a couple of years, and so to the extent that their number of days of attendance in the program changed from one year to the next, and we had their outcomes in those 2 years, we could ask if the days went up, did their outcomes improve?

But we were clearly resting on the presumption that the reason the days went up was not this kind of voluntary "I like this a lot because that drags in all the unobservables" It was that something happened where you just had to be there more or less, whichever.

But it's the identifying information about how endogenous this really was that is lacking. To some respect, I think that the data collection strategies need to be there at the outset to recognize this, and that's, you know, if anything methodological and geeky were to come out of this, you have to think about the identifying strategies for these things in your primary data collection strategies and just imagine that some of these pieces of information are going to be used as part of more of an instrumental variables way of thinking about the world.

This is not necessarily so easy to explain to OMB when they're asking why is this thing here, and you're saying, well, we're going to use it to identify endogeneity. That won't necessarily wash. But I think we could make, I think we can make the argument that the experiment is stronger if we're also able to incorporate kind of a nonexperimental way of grappling with what are actually important questions.

DR. HOLLISTER: Jeff, in supportive work, we had plan variation. We fought like hell to get a plan variation. We got one which said they could stay in the program for a year instead of six months so we let the programs decide if they wanted to use that. None of them did it. None of them took it up. So you can lead a horse—well, anyway.

Yeah. Sorry, Larry.

DR. NEUMAN: This is a little high for me. I want to—my name is Susan Neuman. I'm at University of Michigan.

These are nongeeky questions. But one of the things that I was struck with, especially two of the studies have had really tremendous policy implications. People are saying vouchers versus no. People are saying afterschool programs versus no. People are saying technology versus no.

And so much of what you're doing has very, very strong policy implications immediate, and you see how it's played in the press, and it really is played as a zero sum game. And some of these people today in this group have suggested looking within the black box, and you basically suggested, and I'm not blaming you for it, but you basically said, well, you know, longer day may be a little bit better; such and such may be a little bit better.

And what I would like to ask for is a slightly different kind of study being done in the future. I think we really have to look in that black box. I think we have to understand how many of these kids have had AP courses or not? How many kids really have had a longer treatment versus not?

And by continuing to ask these very, very blunt questions, we are losing so much information, and we're not providing, we're not going any further. We're just continuing the sort of black and white kind of debate. So I'm not questioning what you can do and what you can't do. But I am suggesting that I think the conversations this morning in both the morning plenary and the afternoon are beginning to tell us that we need to look much more deeper in these questions, and really address the key variable.

You were calling it work, but there are lots of other definitions for what I'm talking about.

DR. WOLF: If I can just briefly respond, again, I really sympathize with the motivation for your question. My colleagues and I are continuing to think about and investigate possible ways that we could factor in dosage variation and sort of the concrete manifestation of the treatment into an analysis without risking or surrendering significant self-selection bias so we could get a sort of a more in-depth and detailed understanding of the mechanisms involved.

So I mean we do understand that, and we're trying, we're considering certain alternatives in that area, but at this point we're not 100 percent certain we've found that "Holy Grail."

MR. CORRIN: I mean I guess I would add that for our study on adolescent literacy, you know, I think we've tried to pursue and had some success pursuing certain kinds of nonexperimental, but what we think would be policy relevant investigations within that work that comes up in our reports; and it's not extensive and I don't think it gets to sort of the level of detail that you're pointing to.

I mean one example that I could give potentially is one thing that we think we really know about these programs is that they provided something that was notably different than what was available for the kids in the control groups. We know that those kids in the control groups were not in the vast majority of cases, were not in some other competitive reading class. We know some things about the reading instruction we think they were getting in other classes and how it didn't appear as intensive or feel as intensive as these programs.

We know stuff about their attendance in those classes. It was actually surprisingly pretty good. So there's some other things that we know, not all that are presented here, but I think all of us who do these studies try to incorporate where we can that kind of stuff.

I think it's really important that at the outset as much of that be kind of defined and figured out as possible, and then I think the other thing, and I think the places like the RELs actually may allow for some of this, that you try to launch complementary studies that may drive at some of the same questions, but they expand your capacity to look at different kinds of plan variation across studies so that you build a body of work that allows you to reflect and get at some of the kinds of policy details you're interested in.

DR. DYNARSKI: Yes. Could I just comment a little bit? Susan, I think there's another dimension to your question which is about media management because they treat everything as a blood sport when these studies come out and somebody gets killed.

And they usually then go ask the person who got killed what did you think about that study? And not surprisingly, that person really feels like it's inadequate or inferior in some way. And this really doesn't advance the discourse very much.

As a substitute, I would offer the kind of tradition in health research where a major study comes out and it's accompanied by an FAQ. The FAQ does something like why was this an important question? What did we learn from this study? What other kinds of research is going on that may complement what we learned here?

And so, the effort to put the study into a context allows the public to come away not feeling like something should or shouldn’t happen but rather that science has this kind of sequential ongoing order to it, and we just reached a different place.

But all of us who have been part of that kind of pitched battle around a single study know there’s this much bigger literature and somehow it always gets pushed to the side when a major study is released, but it shouldn’t be because these are accumulations of knowledge.

DR. HOLLISTER: Yeah, Larry.

DR. ORR: I’m Larry Orr, no fixed institutional affiliation.

I want to pile on Mark a little bit here because he had the temerity to actually make a recommendation on the topic of this session, which was that the folks who really should be sorting out the dosage response are the developers, and they should do this in the course of their small-scale efficacy trials instead of leaving it to the large-scale effectiveness trials to do that.

My question for you, Mark, is that given that—well, as I looked at the charts you put on the screen, and I look at the charts in your report, it seems to me that the major problem that you had in sorting out or one of the major problems that you had in sorting out dosage effects in your data was that there was so much noise going on at the level of each of these individual interventions, that when you split that sample in half or more, you really couldn’t distinguish the dosage effect from the noise.

So given that efficacy trials are generally going to have even smaller samples, my question for you is how would you advise the developers to determine and set the dosage effect?

DR. DYNARSKI: Well, in some respect, what we’re really talking about is not the NCEE large-scale evaluation world but the NCER goal world where, you know, it’s goal one, goal two, and so on, and the developers are really, at best, goal two, trying to amass evidence of effectiveness.

And interestingly, in that goal, the statistical significance is given less importance than simply things like effect size and showing proof of concept.

So if you’re going to rely on asterisks to drive your consideration, sure, this is going to be tough, but, okay, that’s clearly not the answer. That’s not the answer you wanted to hear, Larry.

[Laughter.]

DR. ORR: Surely, Mark—

DR. DYNARSKI: Yeah.

DR. ORR: Surely, Mark, you want them to come up with something that is more than just chance sampling variability here. I mean, yeah, they may ignore statistical significance, all the more reason not to trust them with something as important as deciding what the minimum dosage is.

DR. DYNARSKI: Well, but then it would become the government’s role to basically underwrite a commercial sector’s advancement of its own proprietary product. I don’t, I don’t get that one either.

I mean if the issue is that they need to invest adequate amounts of resources to sufficiently power their own studies, then, yes, I will go with that. But, you know, I don’t see why this falls to Congress to underwrite a study of products, which are actually already being sold based on lesser amounts of evidence than what I'm suggesting, what I had the temerity to suggest.

So did I understand your question right, Larry?

DR. ORR: Yes.

DR. DYNARSKI: Basically you're saying why—we didn't have stars in the first place at the big scale, how are they going to get stars at the small scale; is that it?

But I'm saying then why are we seeing these products at all then? If nobody has any evidence that they did anything at any level, then, I mean they're in thousands and thousands—it's a billion dollar industry. So I'll stop there.

DR. HOLLISTER: Okay. I think we're at the end of our time.

[Laughter and applause.]

DR. HOLLISTER: And Jeff, I've got nine questions of the geeky type that I gave these guys.

DR. WOLF: And Jeff, you want geeky, read our study.

[Whereupon, at 2:50 p.m., the panel session concluded.]