STEM EDUCATION RESEARCH DESIGNS:
Conceptual and Practical Considerations for Planning Experimental Studies

Workshop Theme
FRAMING THE KEY ISSUES

Friday, October 12, 2007
Westin Arlington Gateway
Arlington, VA

NSF Workshop on Causal Inference:
How Randomized Field Trials Help Democratize Knowledge in Education

James S. Kim, Ed.D., Assistant Professor
Harvard Graduate School of Education

Introduced by Michael E. Martinez, Professor
University of California, Irvine

MR. MARTINEZ: James Kim is a former history teacher in an ethnically diverse middle school here in Virginia, a graduate of the University of Virginia, by the way, bachelor's degree, and Harvard Ed.D.

Prior to joining the faculty at the Harvard Graduate School of Education, he was an assistant professor and my colleague at the University of California Irvine, and a research associate at the Harvard Civil Rights Project.

Dr. Kim's research interests include the use of quantitative methods to assess the effectiveness of compensatory educational policies for disadvantaged students and the impact of reading programs on adolescent literacy.
He's actually doing the kind of research that we're interested in in multiple sites. He is concentrating in particular on randomized experiments of voluntary summer reading, middle school literacy, and teacher professional development in Title I schools, and, Jimmy, thanks for being here today.

MR. KIM: Great. It's a privilege and pleasure to be here with you. When Mike said elaborating on abstract ideas later, there will be no abstract ideas in my talk because it's going to be incredibly pragmatic and concrete. In fact, I have a pretty simple goal in my presentation, and that is to share with you some really practical lessons that I've learned in planning and implementing randomized field trials in schools.

[slide 2] My substantive area of expertise, as Mike mentioned, is experimental evaluations of reading interventions, but what I want to stress is that the political and the administrative and methodological issues are really the same regardless of your substantive area of expertise so it could be math, it could be science, technology, literacy, the task of undertaking a randomized field trial has very similar issues across those areas.

And the way I would like to communicate to you those pragmatic practical lessons that I've learned is to talk to you about a voluntary summer reading intervention that I've been working on for about ten years where randomized field trials are central to this program of research that I've been working on.

And I have four lessons that I want to distill and extend upon, and ideally at the end of my presentation, if you have just one or two lessons that you can
Okay. So let's start, first of all, with nothing related to randomized field trial but some basic problems in our society [slide 3]. As Mike mentioned, I was a middle school history teacher, and when I taught in a very multi-racial middle school, I observed two facts about my students. And I assume many of you are former classroom teachers or you work in schools, and so you have a lot of anecdotal observations.

Well, here are the two that I made that are very significant. One is when my children entered seventh grade in September, they were all over the map in terms of their reading levels, their knowledge of history, but by June, I felt like they had all learned. They had all grown over the nine months' school year, and this strikes me as pretty good news. Children should learn while they are in school and being instructed.

But the second fact that I observed is that during the summer some kids really learned a lot. They went to Gettysburg and learned about the Civil War. They went to the Smithsonian. They read all the Harry Potter books. Other students of mine, they did nothing, watched TV, and consequently they seemed to lose ground in their reading skills as well as their knowledge of history.

Well, I wanted to put some empirical meat on these anecdotal observations I was making in the classroom, and I came across a study by Barbara Heyns, who is a distinguished sociologist, who in 1978 did a longitudinal study looking at summer reading in the Atlanta public schools.
I just want to highlight a couple of points in this figure [slide 4]. I think you can see it fairly well. Vertical axis is achievement measured in grade equivalence on a standardized test of vocabulary.

The horizontal axis is time so Spring '71 to Fall is how much kids learn in the summer. Fall to Spring is how much kids learn during the school year, pretty straightforward.

The dash lines are White students. The solid lines represent Black students, and here's the big finding. The big finding is if we look at what happens from the Spring to the fall, the first set of lines, you'll notice that all those lines are uneven. You'll notice that in general the reading gains are larger for White students and for middle and upper income students; for Black students and lower income students, the reading gains are either flat or they go down.

But it look at the next set of lines from Fall to Spring, you'll see that all of those lines have kind of this upward slope and, in fact, statistically, those are similar, and so we could conclude that there's a similar rate of growth for all students during the school year. Okay.

So what are we learning here? Kids seem to learn about the same during the school year, and then in the summer, there are all these kind of inequalities that seem to form.

Well, Heyns went one step further, and she gave a survey to the parents of these students in the fall, and she asked the parents, you know, what did
your kids do over the summer, what kind of activities? She had a ton of
different activities, you know, whether they went to the library, whether they
went to summer school?

But here's what she found [slide 5]. She found that the single summer
activity that is most strongly and consistently related to summer learning is
reading; it was the only measure that was related to reading gains over the
summer.

But it went one step further. What she found is that the number of books
read during the summer is consistently related to achieving gains. The strength
of this relationship often exceeds that of socioeconomic status when prior
achievement is controlled. So conditioning on prior test scores, the beta
coefficient for the books read is larger than family income.

Okay. Now why might this be important for me as a researcher? Because
I as a researcher can't do much about children's family income. I can't do much
about that, but I can certainly think about how I might be able to manipulate
the number of books children get or how much they read over the summer.

Well, it turns out that these findings have been replicated in studies in
Baltimore, New Haven, Dallas, nationally representative samples from the '70s,
'80s and '90s in Title I studies. So this is a pretty common phenomenon.

Well, more recently, there was an analysis conducted on the Early Childhood Longitudinal Survey Kindergarten Cohort of 1998 [slide 6]. It's a
nationally representative sample of about 20,000 kindergartners who have been
followed now for several years, and Roland Fryer and Steven Levitt, two economists, wanted to answer the question: “What explains the racial ethnic disparities in reading and math in the Fall of kindergarten? “

Now this is before kids ever start school, and they wanted to see if there are certain background variables that could explain these racial ethnic differences, and I think this finding also is related to Heyns, and you'll see in a minute.

Column six, what that basically shows is if you look at a coefficient for Black and Hispanic, Black and Hispanic children are about .4 of a standard deviation behind in reading relative to White students.

What happens when you control for SES? It goes down. Once you control for the number of children's books in those kids' homes, the Black/White gap is zero. It explains the entire disparity between Black and White students. Hispanic/White gap is about a tenth of a standard deviation.

So why am I showing you this? It's because again correlational data suggests something important about books at home; how much kids read.

Well, here's my first lesson that I want to share with you in thinking about randomized field trials [slide 7]. Correlations provide very important clues about variables worth manipulating in educational experiments. Mike said, look, randomized field trials are extremely expensive; we can't subject every intervention to field trials.
So we got to be really careful about what is it, what's that one variable that we might want to manipulate? And, in fact, both Heyns and Fryer and Levitt alluded to this. Heyns said in terms of a policy recommendation, increasing access to books and encouraging reading may well have a substantial impact on achievement.

Fryer and Levitt said if we include the number of books in our regression model, it reduces the Black/White gap on math to less than a fourth of a standard deviation and completely eliminates the gap in reading. The gap for Hispanics also shrinks.

All right. Let's do a little thought experiment here [slide 8]. This is the late 1990s. I'm very intrigued by this, and the question I have is where do we go from here? So I might be tempted to think two things: I might be tempted to think, number one, that summer reading loss happens everywhere; it's just a widespread phenomenon. It must, part of the achievement gap must originate outside of school in the summer.

Okay. The second tempting thought that I might have is that if we give kids books, it will cause them to read more and improve their comprehension, and wouldn't that be great. I mean talk about a cost effective intervention, that would be it.

Well, I want you to do something at this moment. You might be thinking about doing field trials in your own work. What I tried to do at this point was to put myself in the shoes of the school district administrators whom I would
approach to do an experimental study, and try to anticipate the question, and here's the question that I anticipated and, in fact, they asked me this question.

Of course, I have the benefit of hindsight today so maybe I made this up, but it actually happened. Well, here's what they said. They said, you know, Jimmy, before you randomize our kids to any kind of summer reading program, I want to know whether the problem of summer reading loss exists in my local context.

All right. So all of us when we think about field trials, we have to do our experiment in one or two school districts, but we've done all of this literature reviewing saying, you know, these variables are important, but often what we'll get from a district is, well, what about our district and what do you know about our district, and so here's what I did.

I didn't do an experiment. I did a series of correlational studies on summer reading which were the precursors to my randomized field trials [slide 9]. So from 1998 to 2000, I undertook a series of analyses of summer reading loss just to see if this problem exists in this district, and what I paid careful attention to was the political and administrative context.

So let me tell you what was going on in this district politically. Politically, there had been huge demographic changes in the enrollments so that minority students pretty soon were going to be the numerical majority, and so the district, the school board and the superintendent, made as one of their top five priorities the reduction of racial ethnic disparities in reading and math over the next five to seven years.
So I knew that any study that might address that would be a winner. Administratively, I knew that I had to build stakeholder support for this work. All right. So I needed to get the director of research and evaluation kind of on my side to be my sponsor. And what I asked her for was the following, I said, look, you guys test your kids in the Spring and the Fall, third grade- fourth grade, fifth grade-sixth grade.

There is all this archival data. It's student level data. Very sensitive. If you give me that data, I'll do the analysis for you. And they agreed to it. So without going into all the models and what the regression analyses were, the simple finding was this: we basically found that minority kids, low income children when they transition from third to fourth grade, fifth to sixth grade, in those summer months, they lost ground in reading relative to White students and middle income students.

Okay. So big deal. How is that related to an experiment? Well, it's not yet. What it does do is it prompts kind of the next logical question, which was if that's happening, then what are the kinds of summer activities that are related to summer reading gain? So, again, no experiment. We're still with correlational data.

So here comes study number two [slide 20]. We know that large survey studies provide more correlational evidence about which independent variables are related to which dependent variables. So in the summer of 2002, I surveyed 4,000 children in 18 schools in this district about what they did in the summer, and I really had two goals.
And I think when we do randomized field trials, we have to have kind of these two goals in mind. One is a political goal. The political goal is that we have to continue to deepen stakeholder support.

So what I did is I decided to ask two different individuals to be sponsors of this project: the director of the minority student achievement office; director of Title I. And now what I'm building is I'm building stakeholder buy-in for this work.

And the empirical goal is to isolate the relationship between how much kids read in the summer and how that's related to achievement scores.

Two, it's to look at how much access to books children have by different subgroups. All right. So now think about the question that we might want to answer. The next step in this study, in this series of studies, was would increasing access to books produce significant gains in reading comprehension?

Now, we are talking about an efficacy question that is very suitable to an experiment. So you think you're ready for an experiment. I thought I was ready for an experiment. In fact, I was actually pretty ambitious and I thought why don't I just go back to those 18 schools, and we'll do a huge randomized field trial in 18 schools, 4,000 kids, that would be great.

But here's what happened in the real world [slide 11]. My sponsor said to me, hey, Jimmy, why don't you come back and give these presentations to the principals and the teachers and talk about what you learned, and so I started
giving these presentations like here, and they're all convened. They're very excited because they participated in this huge survey study.

And I was trying to explain the regression results about how if kids read about four to five books over the summer, they do better than kids who read very little, conditioning on prior test scores and including fixed effects for classroom blocks, and so forth and so on. And that minority and low income children appear to have fewer books at home. And a teacher raised his or her hand and said, well, what about motivation? I mean some of my kids hate to read and some of them really love to read; don't you think that matters? And I said, yeah, that probably matters.

Another educator raises his hand and says what about family background? Some of my parents are poor; they can't afford books. Others said, well, some of my parents don't read with their kids. And I said you're right. That might matter as well.

Well, by the end of our discussion, they had listed a series of variables that could potentially explain why we're getting these results, and I put up my hands and I said you got me. Those could all be true. And, of course, we know statistically, there's omitted variable bias. If any of those variables are related to my outcomes, I know my treatment effect is biased.

And what I said to them is I said, well, how can we answer this question more conclusively? And what could we do? And straightforward answer that I said was, look, if we randomly select kids into a program to encourage them to read in the summer, what we're going to do is we can be a little bit more
confident that on average the kids who are getting the program and the kids who are not are similar on all the variables that we've observed as well as the unobservables, and that is the power of an experiment, and everyone understood that. Motivation, failing parents, all those variables that I did not include in the model.

So here's the key decision that I'm facing that you might face if you're thinking about how to plan a randomized field trial. So of those 18 schools, where could I do a pilot experiment? I pretty much decided at this point that doing an experiment in 18 schools would be unwise. I had a lot to learn about the barriers we would face.

And here's the strategies that I took. I basically wrote two-page summaries of this research in very kind of lay-friendly language and I sent it to all the principals and all the teachers in those 18 schools. And I did nothing. I just waited. One month goes by. Nothing happens. Two months goes by. Nothing happens. Three months. We're at four months. And finally I get an e-mail from one of the principals, and she starts querying me about the study.

She says this is very interesting, Jimmy. What's going on here? Well, if it's books, can we get more books to our kids? Do you want to sit down and talk about a study we could do? Of course, to get an invitation from an educator to do a study is what we dream about, right, as researchers.

And, in fact, as we talked, I invited her to participate in a pilot. So here's lesson two about randomized field trials [slide 12]. The process of doing pilot experiments builds capacity for randomized field trials.
Now, I know that term "builds capacity," we throw that around a lot so I want to flesh that out a little. I want to explain to you what does building capacity mean? What does it look like?

Well, Mary, the principal, was the principal of a very large elementary school, and we decided to conduct an experiment where children in grades one through five would be randomly assigned to get ten books in the summer matched to their reading levels and preferences or in the Fall after post-testing.

And why was it important for us to have grades one through five? It's because, as many of you may have heard, Jeanne Chall pointed out often, that in grades one through three, what do children learn? They learn to read. In grades four and beyond, they use reading as a tool for learning the new; right.

So young children, if they don't get a lot of support, they might have more difficulty reading books on their own, and in the scientific literature, among psychologists, this is actually a very important question, with grades one through five, we not only had more power, but we could also answer this question about interactions possibly across grades.

So, look, one thing I want to tell you is sometimes it might sound like in my talk, Jimmy, you know, you're constantly bowing down to the educators; you're doing whatever they want. That's not true. We've got to articulate questions that are significant in the research literature, kind of collectively, jointly.
And so that was a question that we asked and it led to kind of the design, but here's what capacity building looks like. I said to Mary, Mary, you and I are teachers; you and I are former teachers. We know that anything in education, it succeeds or fails depending on whether teachers buy in and implement it, you know, whether they believe that this is worth doing.

So I want you to do me a favor and let me talk to all of your elementary language arts teachers, first grade teachers, second grade, right up to sixth grade, in little meetings. And here's what I said to the teachers, assume that you're the teachers. This is what I said, I said, you know, first of all, I want to define what success looks like in the study. Success is a scientifically first-rate implementation; it is not positive results.

You know, I can't control—why would we do the experiment if we knew what the experiments were going to be; right? It's no point in doing it. But it is not controlling the results. It's having a first-rate scientific implementation. And then I said to them, you and I, if we work together, we can have a very high quality study, and here's what we need to get this high quality study.

One is we need really high participation rates among your kids. Most of the districts that we work in require active consent, and this district did, so we had to get permission forms out, and I said to them I can't have 50 percent; I can't have 60 percent. Can't have 70. You know we want about 95 percent, which is pretty unheard of, but I think we can do it.
So what we did was we had developed all kind of incentives for the kids. We told them if you turn in the forms, some teachers offered ice cream, some gave them popcorn parties, free recess. Some did all of those things.

And then we had three backup plans on those consent forms. I asked the principal and the teachers to call back those parents who didn't return the forms. Well, we got really good participation rates.

Why is that important? It's not just a nice story. It's a very important point about external validity. If you have 60 percent of your kids participating and they happen to be the kids who are all from high SES families, and you've excluded 40 percent who don't return the forms, you can't make generalizations to your entire sample; right. So this is actually scientifically very important.

The second thing I told the teachers is the teachers said to me, oh, I'd love to know which of the kids are going to get the books in the summer, and then I can kind of help them out a little. And I said, no, can't do that. You're going to be blinded to the experimental condition of your children. Randomization is going to happen after the pretesting; it's going to happen after you go home for summer vacation. You will not know. We can't have your expectations be the reason why we get any outcomes. So it's going to be a blinded study.

And the third thing I told them was we really need complete administration of pretest and post-test. We need the pretest to match the books to the kids, we need it for our analysis to improve the precision of our
treatment effects, and we obviously need the post-test if we don't know what happened with the intervention.

All right. So you might be sitting there going, wow, Jimmy, that sounds great. You must have done this experiment. It's a nice story. You found positive effects. Well, here are the results of the study.

In fact, I found no significant impact, none. But what I did find is that all of the positive effects when we just looked descriptively at effect sizes, they're all concentrated in grades three through five, all with the older kids.

Okay. So theoretically this is making a lot of sense. Practically, we can't do much with it; right? Because we can't reject the null hypothesis at this point.

But I'll tell you what was very important about this study and why capacity building is so important in randomized field trials. One is that, you know, that principal Mary who came to me, she's now my colleague and teammate. And what that means is it is not the researcher who's going to do research to the schools, but it's the researcher and the educator who are going to do research with the schools. Right.

So she's now my teammate, and I'll tell you what she did in setting up the field trials in a minute.

The third thing that we've done is we've deepened capacity for randomized field trials with teachers and principals. Well, guess what? You know when we now have that huge training with teachers and we have to
explain to them randomization, all the tests that they give, and they start objecting and start having questions, I'm not the only one who's going to answer those questions. I'm going to ask my teacher colleagues who have been through this to help me answer those questions.

Why? There's always more credibility when your colleague in the same profession kind of can tell you, hey, this is what it looks like. Okay.

And here's the final and perhaps just as important reason why the pilot is ex--we now have good data for a power analysis [slide 13]. Randomized field trials are very expensive. You need to figure out how many kids, how many teachers, how many classrooms, schools you can support, and here's what we learned from our pilot.

Any study that you do, you want to do a power analysis. Basically largest impacts are in grades three through five. The effect sizes are between about a tenth and a fifth of a standard deviation, very common in education, but the sample sizes are very small.

What's the "n" here? It's actually an "n" of one. It's one school. How much can you generalize beyond one school?

The other problem is this district cares a lot about NCLB subgroups, minority kids, low income children. They're very concerned about the achievement gap, as I mentioned to you earlier, but our sample sizes are far too small to estimate separate subgroup impacts.
Well, here's what the power analysis suggested. If we assume an effect size of about .15, we assume that the baseline covariate, which is the Spring reading score, we gave a Stanford 10, it's a very widely used standardized test of achievement. It's correlated about .80, .85 with the post-test, and so we figured that we need about 422 kids to have an 80 percent chance of finding a significant difference on the post-test reading outcome.

But if we factor in attrition over the summer, we need a lot more. We need more than 422. We need more in the ballpark of five to 600 students. Now, you can see why this is really important.

If I don't do this pilot, I have no idea how many kids I need in my study to have a chance to detect fairly small effects that we usually find in education. This is not only true in reading and the work that you do in math and science; it's also probably true—we just don't find gigantic effect sizes.

Okay. So I haven't even talked to you about randomized field trials. Now, I'm going to talk to you about doing experiments in multiple classrooms and school settings that involve over a thousand kids.

So what happens between the end of my pilot and the start of the randomized field trial in 2005 [slide 14]? What I did is I formed what I call a design team. You might have heard of something called a Strategic Education Research Enterprise, SERP, developed by the National Research Council.

The goal is to have researchers and practitioners work together around questions of importance to the educators.
Well, I thought to myself, look, that's a good model. It's a good model because if educators work with researchers to pose significant questions, they have incentive to pay attention to the answers to those questions and they have an incentive to implement those research designs very well.

So I formed what I called a design team and the design team included the principal, Mary, two of her top reading teachers, one program evaluation specialist, and I'll tell you what I asked the design team to do, and this is actually a pretty practical suggestion. When you think about field trials, one thing we as researchers do not have often is what I'm going to explain to you, what these folks were able to do.

One is the principal's role to invite her colleagues to be involved in a randomized field trial. We as researchers are not principals and teachers. We don't, we simply don't have as much credibility with other principals and teachers so we need allies in this process.

And Mary's role essentially was to invite her colleagues to be a part of a larger experiment. The teacher's role was to develop lessons. What we learned from the pilot was that, look, if you just give books to children, that's probably not going to be as effective as providing some sort of scaffolding and instruction at the end of the school year to teach the children what to do once they got those books.

And so we looked at the National Reading Panel report as kind of a base of scientific literature, and the teachers and I developed a series of lessons that
all teachers in the field trial would teach in the last month of school to encourage their children to practice certain reading skills.

Why is the teacher's role important? Well, they're the ones who are going to do all the training for me. Right. So this is very important. In a lot of educational interventions people will often say, well, your intervention worked because you had a Ph.D. professor from some top university do the training.

We want teachers to be teaching teachers because we want this intervention to be scalable. Why do we do randomized field trials, at least on a kind of pragmatic level? It's because we want to see if we can bring the intervention to scale; right? And this is a good test of it.

If teachers can teach teachers, and we can kind of replicate these results, it gives a sense of scalability.

What is the program evaluation specialist's role? What's one of the toughest tasks that we have as researchers when we do randomized field trials? It's getting access to student level data. That is really tough. And so the program evaluation specialist was also the test coordinator, and his job was to be the keeper of the data.

I would go back to him to ask for all the student level data on the demographics, pretest scores, other information, and that person would have to create research IDs and then give me the data file because most districts do not release student names, student original survey IDs, original student IDs.
Now, it doesn't seem like that important of a job, but it's very important. Why is it important? Because as in Tennessee's Project STAR experiment, if you want to get data on those kids 20 years later, you need to have someone who can link up those kids' original IDs to all the outcomes that you're interested in.

So this is my core team. These are the folks that I've asked to be my allies as we prepare for the large randomized field trial, and here's how the field trial evolved.

[Slide 15] We approached the superintendent who is in charge of these schools where Mary was a principal and what she agreed to was to allow Mary and I to give presentations to the principals in her group of schools to invite them to participate.

And guess what happened? Well, we got a lot of objections. Right. I mean Mary is great. She really supports this work. She's my colleague, my teammate. A lot of objections. What are the objections you think you're going to hear when you try to do these large experiments in schools today in our context?

Well, one objection went something like this: *Are you kidding me? More testing in the NCLB era? You know we test our kids all the time, and you're going to ask to give a test at the end of the school year and another test at the beginning of the school year?*
Second objection. Randomly assign kids. What are you talking about? Why don't you just give it to all the kids? I mean this sounds like a really good deal. Why the heck would you want to withhold this from the students?

Third objection. There have to be hidden costs. You know you researchers, you tell us you're going to give two tests that take 30 minutes, and then you really give four tests, and they take like five hours. So there have got to be these hidden costs; right.

So we're getting objection after objection, and so Mary and I, we knew kind of going into this we'd have to think about some of these issues. Some of them we had thought about. Some of them we had not. And here's what we did, just practically. We don't have a lot of options other than to kind of be straightforward with them and try to minimize the costs, right, because we're pretty clear that we think the benefits to the students is pretty high. But we're going to now try to minimize the cost. Here's how we try to minimize the costs.

One is we said the Iowa Test of Basic Skills, it's about a one-and-a-half hour reading test—we're going to give this after NCLB test. It turns out that most school districts in the United States, they give those end-of-year No Child Left Behind test anywhere from February to about May. And some states like New York do it very early. Other states do the testing late.

So there's actually about three or four weeks at the end of the school year where kids are not being tested, and the teachers are very tired, and so what we said was it's a one-and-a-half hour test. It's group administered, and
what you can think of it as is it's just an activity for your reading block, your English language arts block, and then we said, by the way, you're going to get these really terrific lesson plans developed by teachers that take about a week to do.

You've got three weeks at the end of the school year to kind of fit them in in your language arts block, but think of it this way, guys. We're giving you a reading test that takes about an hour and a half. We're giving you free lesson plans and that means you can cut down on the amount of planning time you have to spend to develop lessons at the end of the school year when you're really tired.

Okay. And you can reuse these lessons next year. Right. So if you do it this year, you can reuse it next year. And then, by the way, the test that we give in the Fall, if you give the test in September, we will turn around those tests and give you the information on how your students perform relative to a national norm group in early October, and this district at this time, in 2005, it actually did not give a lot of formative assessments.

Kids got grades from their teachers, but they didn't give standardized tests to identify where they were relative to a normed group. They just have these end-of-year NCLB tests, and they knew they needed to get the kids ready for this.

So what we said was, look, if you get this test data in the Fall, you can identify which of your children are well below the norm on a nationally normed test like the ITBS, and they all saw the value of that because one of the
concerns that they had was they had all these intervention strategies, but weren't always sure which kids really needed those interventions, and so we said, look, we'll give you the information and we'll customize it for you.

The next thing that we did was we developed an MOU, a memorandum of understanding. When we do randomized field trials, one of the things that's pretty important is--I'm going to come to this theme later, but it's not just enough to have an agreement with the superintendent. It's not just enough to explain to the principals what you're going--you actually really need to have some paper that the teachers read and that they understand what the study entails, benefits and costs, and that they sign.

And so we had an MOU that outlined all the benefits and all the costs of participating in this study, and what we did with the cost is we actually included more time on every activity. The ITBS takes an hour-and-a-half for the students to take the test, but what we said was plan on two hours.

When you think about your studies, when you calculate the time it takes to administer instruments, you always have to think about the preparatory time the teachers are going to take, both to understand the instrument and how to give it because instruments are given differently. Some are one-on-one. Some are large groups. Some involve accommodation. Some have exemptions for kids who are special. There's a lot of little details that you've got to cover.

You've got to think about that preparatory time. Then you've got to think about the administration time and then you've got to think about the time of getting those tests back. So we have this MOU, and it's May of 2005.
I had enough money in my research funds to support about seven or eight
schools, and what we told the principals was, okay, is you're going to go back
to your schools, you're going to explain the study, you're going to invite your
teachers, and then if you want to do this, we want to see the signatures, and it's
going to the superintendent to say that you want to do the study.

So we had 14 schools that said they wanted to do this study. Okay. So
obviously most of them saw the benefits of this and we really conveyed to them
that the benefit-cost ratio should be about two to one. Okay. So twice as much
benefit as cost.

So at this point, if you think about where we're at, it's 2005. I started
this work in 1998, and we had built a lot of stakeholder support, and to me as a
researcher, the real test of whether I've built capacity or buy-in to this
research, it's money. Right. You know districts can say, oh, this is a great
idea, we'll do this study, it's great, but one criteria for us as researchers to
determine whether they're really going to do this well and this is going to be a
scientifically first-rate study is are they willing to put in in-kind contribution?

So I asked the deputy, the Title I director, said, look, if you pitch in
$20,000, we can include about a thousand students and we can have 12 schools,
and we can have a larger sample. That's certainly going to help us from a
research side.

Now, here's a key question that you might face as a researcher. What do
you think we should do? Should we conduct one large randomized field trial in
these 12 schools or should we have smaller planned variation studies?
When I say randomized field trial, I'm assuming that what comes to mind for most of you is something like Tennessee's Project STAR that involved over 79 schools, eventually over 10,000 kids in the state of Tennessee. One large experiment done over a period of time to answer a question about efficacy, the efficacy of class size reduction.

Or you might be thinking about what Congress recently funded, a study published by IES, was conducted by Mathematica, on technology programs, over $10 million, do technology programs to improve student outcomes? Results weren't great, but it was a very large field trial, lots of districts, lots of schools, kind of one main question; right. So we could do that. We could have smaller studies.

Here's a really pivotal point at which I think I learned an important lesson. I think to some extent the answer depends on the stakeholder question. So here's lesson three.

Lesson three is this [slide 16]. Planning two randomized field trials I think helps address multiple questions asked by stakeholders and indicates whether your findings are replicable. All right.

So if you've got questions that are of deep concern, not only to the superintendant, but to the principals and the teachers and the language arts person and Title I, you've now just got buy-in from district officials who care about the answer to the question, who care that we implement the study well.
But what's equally important is if you have two experiments going on simultaneously, you have a sense of whether your findings, your intervention is replicable, whether the results are replicable, and I think replication is very important in education because we often find that, look, you may get a result one time in one context, but you do it again, you don't find the results again.

So we care a lot about replication and understanding the efficacy of our interventions. So here are the two questions that actually emerge that we jointly articulated as a team, as well as yours truly as a researcher thinking about what is an important question in the scientific literature, because to be really honest, in our line of work, things have to be publishable.

So we've got to think about the broader scientific literature while at the same time we're thinking about what do practitioners care about? Well, here's the first question, and this leads to the first randomized field trial.

The first question was can a voluntary summer reading intervention improve reading achievement among different subgroups of students in a diverse sample of schools?

Basically, the superintendents and principal said to me, look, Jimmy, our biggest concern right now is making AYP. We are failing AYP because of the NCLB subgroups. Anything that improves those kids' outcomes is going to be a value to us. So ideally, whatever you need to do in your sampling design, we want to know what are the effects for minority students, for children who are less fluent readers, for those subgroups.
Okay. So that was the first question we articulated. Now, I know as a researcher my job was to make sure given the power analysis and all the work we had done, when we sampled, we needed to make sure that we had those subgroups in our sample.

Well, it turns out that the teachers and the program evaluation people, they asked a similar question, but a little different. What they wanted to know is if those teachers provide lessons, if there's skill instruction, and the access to books in summers, does that lead to greater reading gains? Is it the instruction plus the books or is it just the books? Okay.

And the question that I kind of put in there as a replication question, are the treatment effects similar to the main study? All right.

So here's what this field trial looked like [slide 17], and now what I'm going to do is kind of explain to you what we did and what we found.

The 2005 Project READS field trial was an intervention evaluation, and the intervention had three parts. One is all the children received—the children in the treatment group are receiving eight books over the summer that are matched to their reading preferences and reading level.

We used something called the Lexile Framework which puts reader ability and text difficulty on the same scale. It uses IRT modeling. It's just a one parameter Rasch model, and so that helps us kind of match the books to the kids' reading levels, and that's the first part of the intervention.
The second part of the intervention is that the teachers taught children how to read aloud. Many fourth graders, children who are making the transition from learning to read to reading to learn, they are not fluent in their reading. They stumble when they read aloud, and if you struggle too much with decoding, it's very hard to focus on the meaning of text. So we added lessons on fluency and reading aloud with your family member.

The third thing that teachers did was they taught these students how to use comprehension strategies when they didn't understand what they were reading by themselves. Look, all of us in the room, when we don't understand something we read, we know what we should do, right? You just go back and reread it.

What else should we do? You know before we start, we probably ask do I know something about this? Is this similar to something else that I read?

Well, these very useful strategies are things that fourth graders don't always know. So the teachers taught the children how to use these strategies. So what is the treatment and control in this randomized field trial that involves--the first one involving ten schools, a little over 500 students. Here's what happened [slide 18]: The treatment group, eight books are mailed to them over the summer. The books are matched to their reading level and reading preferences.

They receive a reading postcard where we ask them questions about their book. The teachers have taught them over two weeks about how to answer
questions on this postcard. That gives us data on implementation, and there's a letter to the parent about reading aloud and reading with your child in English and Spanish.

Okay. So that's kind of the packaged intervention. So what happens to the control group? They get eight books after post-testing. This secures buy-in from the district, the schools and the parents. But it's also kind of important from a research perspective. In my opinion, I think this potentially reduces threats to internal validity.

When we talk about Hawthorne effects in education, what we mean is that if you do something new, it might just produce better outcomes. Well, if you think about the design, if both of the treatment and control group kids are getting the books, they should, in theory, both be equally excited about the new program. There shouldn't be any difference in kind of level of excitement.

I'm also a little bit less concerned about John Henry effects. There's no reason to believe that the control group is all of a sudden going to work harder on the Fall test because they feel like they were deprived of some sort of valuable intervention.

So here's what the sample looked like [slide 19]. In our ten schools, Black students, Latino students, Asian students represent about two-thirds of the sample, and they were actually very even in terms of their numerical composition. Why is this important? It goes back to the research question. We're able to estimate effects on subgroup outcomes.
When we looked at the Iowa Test of Basic Skills in oral reading fluency, it turned out that the sample of kids in our study, the mean was right at the national norm, so the ITBS mean national percentile rank was 52. So these are kids who are neither really, really struggling nor really, really high performing. It's fairly representative of maybe the distribution of the national norm.

Even on motivation and reading attitudes, when we look at comparison to a national norm group, right about average. Okay. This all is important in the inferences we're able to draw on this question of external validity.

So what did the final sample look like? [Slide 20] We had about 552 kids. Remember I told you earlier in the pilot, when we did the pilot analysis, we factored in attrition. In fact, we did have attrition, but there were no significant differences between groups.

The baseline data, we checked for difference. We wanted to make sure random assignment worked. Two groups were statistically equivalent on the pretest. And our analysis is pretty straightforward, just using a regression analysis as a pretest, post-test as the outcome, and then we included various fixed effects for the randomization block or the classroom in which the children were nested.

So what did we find? [Slide 21] Before I tell you what we find, I just want to say a word about effect sizes. You all are probably familiar with the fact that in education the effect size that we often find is about .20 in effective intervention, the treatment group scores about a fifth of a standard deviation higher than the control group.
Small class size reduction, Success for All, the effect of remedial class size, all around .20. Why am I pointing this out? Well, for a couple reasons. One is we want to relate our findings back to the broader literature, and, two, as I said earlier, no matter what your substantive area of research, we don't find huge effects in education. It's very rare.

I've given you some interpretive pieces there. .20 effect size means the average child will move up to about the 58th percentile rank on the distribution, and so keeping this in mind, I'm now going to tell you what we actually found.

Well, what we found in the main study [slide 22] is that all the positive impacts on the ITBS were concentrated among the lower performing subgroups. All right. So we find an effect size of about .08, not that large overall. But the effect sizes are larger for Black students, Latino students, the less fluent readers, the kids who had fewer books at home.

Okay. So certainly a gratifying finding, but now I want to pause. Okay. Suppose, suppose we stopped here, you know what have we learned? Okay. We learned that this kind of packaged intervention seems to help NCLB subgroups, the students who belong the NCLB subgroups, and that's certainly a gratifying finding.

Remember what I told you earlier. We decided to do two field trials, and in the sub-study, the second experiment, we wanted to know which components are driving the reading gains? [Slide 23] Main study, we kind of learned the
Jimmy Kim Cadillac intervention: books, teacher lessons seems to make a difference.

But what we did in the sub-study is we wanted to replicate this effect and we wanted to know in order to answer the replication question, the kids getting the books, fluency comprehension instruction, do they do better than a control? That would answer the replication question.

Do the kids getting the books and fluency and control do better than the kids who got just the books? That would tell us whether instruction matters. And do the kids who just got the books do better than the control? That would tell us if it's the books that's driving the gains. Okay.

[Slide 24] Here's the diagram showing the design. This is modeled after the Tennessee Project STAR experiment. We randomly assigned kids and teachers to four different conditions in two schools in grades three, four and five.

If you start all the way on the far left, one group of teachers and kids, they just get books. The teachers teach the children just how to read books over the summer. That's it. Group two, the teachers are teaching the children how to read the books and to practice fluency. Group three, the fluency and the comprehension strategy. That's what we tested in the main study. And then the control group.

Okay. What did we find? [Slide 25] What we found is if you look at the far left bar, .02, the difference between the kids who got the books and the
control, no significant difference. Right. So just giving the books doesn't seem to have any impact on achievement.

Books and fluency, it gets you a little bit more, but again, no significant difference relative to the controls, and here are the two important findings that we got. The .12 effect size, that's a difference between the full-blown intervention and the control group and the effect size is right in line with the effects we got from the main study.

So it's kind of replicating the main study effects. And the kids in the books, fluency, comprehension do marginally significantly better than the kids who just got the books. So the kids who are getting that instruction are doing better than the children who just got the books.

Okay. Now I want to kind of move to the conclusion in my fourth lesson. Where do you think we're at in the research right now after doing these randomized field trials? There's a problem. It's a problem of an "n" of one. Right.

[Slide 26] I've done this study in one district, and Fred Mosteller and John Gilbert in a chapter in a book edited by Moynihan and Mosteller that was a reanalysis of the Coleman Report, they were reflecting on the state of education research, and how in education we only have like little experiments in single districts, and they said, referring to class size, the final blow may be that all the schools are from one school system in one county in one state in one region of the county, and so from many points of view, such experiments
are based on samples of one, a very small sample for choosing a national program.

The point is this is why we do multi-district, multi-state randomized field trials. We have to kind of see if these effects are generalizable. So here's what happened. Lesson number four [slide 27]: to implement a good randomized field trial, we need the support from teachers, not just a letter of support from a superintendent.

So what we decided to do was, we decided to do a multidistrict field trial Project READS to address the replication question across districts, and just this past Spring, the study involved three districts, a district in Virginia, Massachusetts and California, and in summer of '07 we added California to the study.

It's a planned variation study where we're trying to replicate and we're also trying to intensify the intervention. Okay. I just told you what the lesson was. Well, here's how I learned this lesson.

I'm in California at the University of California, and now I'm approaching this district to try to do my randomized field trial. I meet with the district administrators, the superintendent, director of Reading First, director of English Language Development, and I share with them the findings, and here's what they said to me.
They said, Jimmy, these results are great. But it's Virginia. We're in California. Two-thirds of our students are English language learners. Over 75 percent receive free lunch. It's a different context.

So here's what they said. They said do you think that you could include a parent training family literacy component where you actually teach the parents how to read with their kids so that they also get support?

The second thing they said, and my answer, of course, was we'll think about it—okay. Then they said to me are you kidding me? Are you kidding me? More testing and you're going to pull out our teachers for training? We test—because we are Reading First schools, we test our kids every week. This is literally going to be about the 75th test that they're going to take.

I said, okay, we'll think about that. But they said, nonetheless, we are definitely going to write a letter of support for a grant that you're going to submit to the William T. Grant Foundation because we're going to do the study, we'll do the random assignment, we'll do everything, you know, and we'll do the family literacy, but we're not sure about the testing and we're not sure about the training.

So I collaborated with a colleague of mine who is a literacy expert who had established really strong partnerships with districts because now you might be saying, you know, Jimmy, look, if you're telling me I need to spend seven years before doing a randomized field trial, you know, I'll be retired or somewhere else by then. Well, I couldn't do that at this point. I had one month; right.
So I basically looked around a university and knew of a scholar who had established really strong partnerships with this district and they exist, I'm telling you, in every school district or university. And so we went through the same cycle of building stakeholder buy-in in about two months.

We gave presentations to the district administrators. Then we gave presentations to the principals. They we talked to the teachers, and here's what happened when we talked to the teachers. We showed the teachers the lesson plans. We showed them the books that their kids would be getting.

We told them everything that we told the Virginia teachers two years earlier, and we told them it would be great if you guys gave the test. You know here are the reasons why. Also, that training is very important. We want consistency, fidelity of implementation across studies, and they said we'll do it.

We'll do it. We used the same strategy that we'd used in '05. Look, I'm not telling you this because this is the outcome we're always going to get, but I am telling you this because lesson four is really important. If a superintendent says yes, we'll do the study, it does not mean the study is going to go well, and moreover, if I had just stopped with the superintendent letter, I would never have met with the teachers, and I think our design would have suffered.

Okay. So here's the conclusion. You're probably thinking, Jimmy, what's the point of the title of your talk? You never talked about democratizing knowledge. Well, now I'm going to talk about it.
Why do we do all this? [Slide 29] I think we do randomized field trials because it makes it easier for all the members of our democracy to discuss what works in education and it requires ongoing collaboration with educators. Look, as researchers, we are experts, but we're also members of this democracy where we can't control the use of our research nor the outcomes of our work.

So the four lessons that I shared with you, one, correlational evidence provides clues about variables to manipulate in an educational experiment. Absolutely valuable to find out which variables are worth testing in a field study.

Two, pilot studies deepen capacity to do randomized field trials. They did two things for me. One, building the design team; and two, the power analysis that was so critical in informing my randomized field trial.

Three, if you plan two randomized field trials, you can address the question about replication, and you also have multiple questions that might be of interest to different stakeholders, and if you have different stake--if we have different stakeholders who are interested in our study, they're more likely to implement it well, and they're more likely to pay attention to the results you get.

And four, if we want to implement really good randomized field trials, we really do need the support of teachers, not just the letter of support from a superintendent.
Daniel Patrick Moynihan said this in his book Maximum Feasible Understanding, which was a study of community action programs during the Great Society, he said the role of social science lies not in the formulation of social policy but in the measurements of its results.

And so that is why I do these randomized field trials I think to a large extent. It helps us measure the results of policy innovation very well and a little bit more precisely than other designs and it helps us communicate these results to the political and educational leaders who might use them to improve policy and practice.

Thank you.

[Applause.]

MR. MARTINEZ: Thanks, Jimmy. I just want to say this is exactly the kind of thing that I was hoping for. We have here ideas, nuggets, well, four of them are listed here, but there are more in the talk that I think are directly relevant to the planning, implementation and analysis of data from, research data that ought to be relevant to project staff and to program managers, as well as the National Science Foundation.

We started a little behind schedule. So let's go for maybe five, six, seven minutes with questions and answers.

MR. YIN: I heard a word early on around the pilot stage that I didn't hear later on, and it was "blind." I wonder if you could comment on how that was dealt with in the actual field trial?
MR. KIM: In the two larger studies that I did?

MR. YIN: Yes.

MR. KIM: Good question. So the point about blinding is really important. You can imagine that the students, if the teachers know which students are in the treatment or control group, they would want to encourage let's say the kids in the treatment group, and then you've all of a sudden changed your intervention.

We essentially followed the same protocol that we followed in the pilot. What we explained to the teachers in the main field trial is just, as we had done in the pilot, a successful study is not a positive result; it's a scientifically first-rate implementation.

As a result of that, we need a complete pre/post testing, and we need to blind you to the experimental condition of the students. And so just as we had done in the pilot, what we did in the main study was that we, in fact, did the randomization and assigned the kids after the school year so the teachers did not know which children were in the treatment control group.

Now, your question is very relevant to the second study because in the second study, what did we do? We randomly assigned teachers and kids to literally different classrooms in these buildings where they got different scripts to follow.

Are we blinding at that point? We're definitely not blinding at the point because at that point I'm in front of you, I've just been randomly assigned as a
teacher, let's say, to control group. I know that you guys are in the control group, and what that meant in the second study is that the script we gave the control group teachers, you can do anything you want, you can show videos. You can read aloud to your kids. You just simply cannot do anything related to Project READS.

We had observers to monitor implementation and so forth. But the teachers who are with those other kids, the books, the books and fluency, books, fluency and comprehension, they, of course, know what is happening to those children in the summer.

So there is no blinding, but that's I think a reason why the whole replicability piece is really important. If we didn't have that first study, but we did have the blinding, I couldn't address that question because we had it and because the effects from both of the experiments are pretty similar in terms of magnitude. I'm less worried that the blinding, the lack of blinding in the second study would have been a problem.

So I think maybe the lesson to take away from that is just having that second experiment where you might have less of a good study with respect to blinding gives you a sense of whether the impacts are still the same.

MS. IRISH: You talked about the variabilities in your study and we've all heard that a quality teacher makes a difference.

MR. KIM: Sure.
MS. IRISH: Did you do any research into the different teachers that implemented this program for you and was there any difference in the student achievement based on those teachers and the quality or the level of their interest or enthusiasm about the program?

MR. KIM: Yeah, that's a great question. Let me answer that kind of retrospectively and prospectively. Looking back, we did it in a very cheap and easy way. We basically didn't have a lot of resources at that time, but we had one rater, my colleague. Remember the core team member, the program evaluation specialist. He and I developed a rubric and we did actually observe the lessons, and we scored them.

Unfortunately, the rubric was very course. It wasn't very sophisticated. Nothing like what IES uses in its Reading First studies that ABT is currently undertaking. But it did give us a preliminary sense of what happened in the classroom, and what my colleague and I basically found was it was essentially very even, and in terms of the even implementation, we actually wanted that.

We wanted a lesson that was pretty easy for any teacher regardless of their skill level to implement. We didn't want this to be something that would require a lot of training. We wanted it to be scalable, so we did have that data. It wasn't great, but there wasn't a lot of variability.

Now, let me tell you prospectively, the study that we're doing right now, we're actually in the Virginia site, there are over a thousand students in over 50 classrooms, and we videotaped about 35, 40 of those lessons, and we have much more extensive coding done by two raters.
To address your specific question, does the quality of those lessons, is that related to how those children then did in the summer in terms of their work? What we're finding in those evaluations is there is quite a bit of variability, and so we do want to look, kind of do those impact estimates, is it related to these scores on the fidelity and quality measures? So that's something that I do think is very important to look at.

Yes.

MR. SUTER: I've been trying to figure out how significant your project is and how to figure out whether it's significant. You reported your effect size--

MR. KIM: Yeah.

MR. SUTER: --of about .1 or .12, less than .2, which you said was the average of all effect sizes if we looked at all studies. But if I get your main argument--

MR. KIM: Sure.

MR. SUTER: --it is that the randomized experiment method involved you as the researcher with the participants, the superintendents, the teachers, maybe the parents--

MR. KIM: Uh-huh, uh-huh.

MR. SUTER: --and therefore the scientific results at the end are more democratic, they're more believable by the people who are affected.
MR. KIM: Uh-huh.

MR. SUTER: So I'm trying to reconcile the size of the effect.

MR. KIM: Uh-huh.

MR. SUTER: Is it too small?

MR. KIM: Right.

MR. SUTER: Or is it about right?

MR. KIM: Uh-huh.

MR. SUTER: Or what would we expect?

MR. KIM: Uh-huh, uh-huh.

MR. SUTER: And I'm thinking even a small effect, even a tiny effect, repeated many, many years, over many, many people--

MR. KIM: Right.

MR. SUTER: --is still likely to eventually lead to improvement.

MR. KIM: Right.

MR. SUTER: Maybe what we're looking for is no negative effect.

MR. KIM: Right.

MR. SUTER: In any effect. And I also wonder about the size of the effect given your outcome measures.
MR. KIM: Uh-huh.

MR. SUTER: Because you said nothing about the reliability and validity of your outcome measures.

MR. KIM: Yes.

MR. SUTER: And those can be absolutely horrible, even full of randomized--

MR. KIM: Sure.

MR. SUTER: --error.

MR. KIM: Sure.

MR. SUTER: So that again a tiny effect might mean something.

MR. KIM: Sure.

MR. SUTER: So how do you know?

MR. KIM: Let me try to answer that in kind of two ways. One is with respect to cost and, two, is with respect to the point you just mentioned regarding the reliability and validity of the instruments that you're using.

I really think that the comparison of effect size just kind of blindly with no regard to costs is really problematic. Let's say you develop an intervention and you get an effect size of .40, but it costs a million dollars every--let's say every standard deviation gain is going to cost you a million dollars.
If someone else does an intervention and it costs half of that, in the real world of districts and schools, that's probably just as important as whether there's a causal link between your intervention and student outcomes.

I didn't talk a lot about this, but the reason why the district wanted to do this is they know they have a huge challenge addressing the achievement gap. They don't have unlimited resources, and they don't have a lot of experimental data on programs that they currently run.

So, one is we got to think about cost effectiveness; right. So is this cost effective relative to other programs? The question about cost effectiveness, though, is relative. I can only answer the question about whether my intervention is cost effective relative to some alternative.

So it turns out what happened in this district--it's part of the study we're running--is they actually have a remedial summer school program, and they have this very nice feature. Bob might talk about this in a moment where kids who they think are going to fail the end-of-year test, they give them free summer school, right. And so there is this very nice kind of regression discontinuity cutoff where if they think or, in fact, if they do score 399 or below on the Virginia SOLs, they're likely to get free summer school.

The kids above that, they don't get free summer school. Well, we can use instrumental variables and various kinds of techniques to get a quasi-experimental estimate of the effect of the remedial summer school program, and we can get an effect of the cost, an estimate of the cost.
I can do the exact same thing for my program. So one of the very important points about what you're raising is all of these discussions about cost and effects are very relative. They're relative to these questions about cost, and so when you and I run programs, we've got to think, okay, what's the effect but also how much did it cost me to get that effect relative to other options?

So I agree with you. I think I wasn't very clear in trying to emphasize that while .20 is very average in education, the cost of Success for All in summer school varies a lot. We have to consider it.

Second point about the instrument, I was actually talking to a psychologist colleague of mine, and he said to me, and I said, yeah, my effects are pretty small, and he looked at me and he said, look, I run lab experiments. We develop our own tests. We never use things like the Iowa Test of Basic Skills standardized tests. If you're getting an effect of .10 on a standardized test like the ITBS, that's a huge deal.

Because think about what we could do. We could make our own instruments that are much more aligned to our intervention, and we would want to do that when we're theory building. But his point was tests like the ITBS, they're intended to draw inferences about broad domains of knowledge, and so given how weak this intervention is, I'm not surprised that the impact was small.

And I'll tell you why I used the ITBS. The reason I used it is some of the best randomized field trials that have been done, Teach for America, the Technology Based Program, they use the ITBS a lot, and I just wanted to use an
outcome that other scholars had been using in field trials so we had some comparability.

But it's one of many different options, and this study we're running right now, we're also measuring motivation and a lot of other important non-cognitive outcomes.

MR. MARTINEZ: Let's take one more question.

MR. REED: Well, my question has to deal with the teacher background.

MR. KIM: Sure.

MR. REED: My wife is a Ph.D. reading specialist that teaches at the junior college, and the problems that she's having there is the fact that 60 to 70 percent of the students entering the junior college are going through remedial reading.

I agree 100 percent with what you're saying. The question which I'm curious about, are the teachers who are in this program, to what extent did they have master's degree in reading? Do you have any idea about that because I'm finding from my wife's background that the teachers that she hires to remediate these students that she has, she's making sure that they all have master's degrees with a reading background.

MR. KIM: Uh-huh.

MR. REED: So do you know anything about that?
MR. KIM: You know I don't actually have information on that, but I can tell you right now that the question you ask is important because I think what underlies that is presumably any intervention that involves a reading intervention is going to be implemented better if the teachers themselves are more well trained.

But that's precisely why randomized field trials in multi-district settings are so important. We know, if you think about northern Virginia, one of the wealthiest regions in our country, the teachers here are so different from the teachers that I'm working with in southern California, one of the five largest urban districts, where I just finished--I was just finishing post-testing, and I was getting rosters of the teachers' names and the kids so we could generate surveys.

And when I was going through that, I'm almost certain that these teachers did not have master's degrees in reading, and I started going through the rosters and I was struck the week before school how many of those classrooms did not have a teacher. All right. So do you know who they hired? Well, they hired basically anyone--they hire uncredentialed teachers, right.

They hired--don't even--it's not even a master's degree at this point. It's just, look, do they have a college degree and sometimes less. Why is that important? I think it's important for this broader question about randomized field trials in multi-district settings.

The reason we have a difficult time understanding whether these effects are replicable is because teachers vary so much in these contexts, and we do ask
teachers ultimately to implement a lot of our interventions. So I don't think a lot of those teachers do, and it will be interesting to see what we find.

MR. MARTINEZ: Thank you very much.

[Applause]