Institute of Education Sciences

Fourth Annual IES Research Conference
Concurrent Panel Session

"Why the Research Community Should Take Notice of Statewide Longitudinal Data Systems"

Tuesday
June 9, 2009

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

Contents

Moderator:

Lee Hoffman, NCES 3

Presenters:

Tate Gould, NCES

"What Data Can We Get From Statewide Longitudinal Data Systems?" 5

Sean Mulvenon
University of Arkansas

"How Are Researchers Using Data From Statewide Longitudinal Data Systems?" 20

Jane Hannaway
Center for the Analysis of Longitudinal Data in Education Research

"How Can Researchers Use Statewide Longitudinal Data Systems to Inform Education Policy?" 43

Q&A 62

Proceedings

MS. HOFFMAN: Good morning. The doors are closed so I take that as indication that we're ready to start. Okay. Welcome to this session. My name is Lee Hoffman. I work with IES and I would like to welcome you to a conversation on why the research community should take notice of Statewide Longitudinal Data Systems. We proposed this topic because there's been tremendous amount of growth in the number and capacity and sophistication of these systems over the last 5 years or so, and a fair amount of it has been fueled by Department of Education grants that have supported the development of state systems and that have included a requirement to make the systems data useful and accessible for research, instruction improvement, running schools.

There are three presenters this morning. They're going to look at different aspects of using data from these Statewide Longitudinal Data Systems. I'd like to introduce each one of them briefly, and then in the order that they'll be talking with you, and then turn the session over to them.

Tate Gould, fairly recently, fairly recently, received his doctorate in education policy from the University of North Carolina. He came to NCES about a year-and-a-half, 2 years ago. He had worked with the Hunt Institute in North Carolina on education policy where he was involved in a couple of different study areas.

Dr. Gould has been the principal lead in the IES, the Institute of Education Sciences, Statewide Longitudinal Data Systems Grants. He was in charge of the recent competition that was just awarded that led to something

more than two dozen State Systems Grants being issued to states.

He's managing the upcoming grant competition that's funded under the American Recovery and Reinvestment Act, which will be large. We don't know how many states will be—how many grants we'll be able to fund or how many applications there will be. They're competitive grants, but it's about a quarter of a billion dollars.

Tate is going to give an overview of the SLDS Grants Program and talk about the content and the access to these that apply to research users.

Sean Mulvenon has a doctorate in measurement statistics and methodological studies from Arizona State University. He is a—you got another one out there—he's a Professor of Education Statistics and Research Methods at the University of Arkansas, and at the University of Arkansas, he established and he directs the National Office for Research on Measurement and Evaluation Systems, or NORMES.

Dr. Mulvenon is going to focus on longitudinal data uses and caveats, particularly in growth model research.

Jane Hannaway's doctorate in education was awarded by Stanford University. She is currently a principal research associate at the Urban Institute here where she directs the Education Policy Center. Dr. Hannaway is also the Director of the Center for Analysis of Longitudinal Data in Education Research, or CALDER.

She's the Director and I think I said Principal Investigator. In her presentation, Dr. Hannaway is going to discuss how data from these Statewide Longitudinal Systems can be used to address issues of education policy.

So, having said that, let me turn this over to Tate.

MR. GOULD: All right.

Thank you for coming, and thanks, Lee.

Again, I'm Tate Gould. I'm with the Statewide Longitudinal Data Systems Grant Program. So I'm going to give an overview of the grant program, a brief history of what we've done so far. I'll also talk about what are the grantee states doing in terms of the data they're collecting as well as the access they're giving to researchers, and then I'll also talk about anecdotal evidence of four states and how they're reaching out to the research community specifically.

The legislative background of our program, 2002, through the Education Science and Reform Act. We gave our first grants out in 2005, so it's a competitive grant process. Only State Education Agencies can apply, and in the first year, we gave out 14 grants to SEAs. They range from 1.5 to six million, at least in the first year; 3-year award project periods.

Second year or second competition in June 2007. We had another competition, with 13 SEAs. We gave out 62 million, and as Lee mentioned, just recently we gave out 27 grants to SEAs for $150 million. So, total, we have 42 State Education Agencies that have grants. We've given out 54 grants.

The goals of the program. The major goals are to improve instruction; to increase the graduates with knowledge and skills for succeeding in the postsecondary and workforce; simplify the processes of using their data in making decisions, reducing the burden of federal reporting

and state reporting; informed decisionmaking; and as well as permit the accurate use of timely data.

So those are the major goals of the program although each state, many states and many grantees will go obviously above and beyond with what they do with the SLDS, but these are the underlying goals of the program.

In terms of our grantee states, and this is also on our website which we'll show you at the end, we have 42 State Education Agencies. DC is one of those, so we have 41 states and DC.

In our last competition, we had several repeat winners, so we had 12 repeat winners that received grants in the first and the second year, and many of the repeat winners when they got their third grant or their second grant in the third competition, they were doing, they're moving beyond the K-12 space and they're linking their data, their early childhood to postsecondary to labor, building research portals. So they're moving beyond the K-12 focus, which was mainly the activity in the first and the second year, is building that K-12 student unit record data system.

A map of our grantee states. Currently we have nine states that have not received grants, but it is worth noting that some of these states have funded these systems on their own, using state funds based on their internal support for this effort.

So just because they have not received a grant does not mean that they don't have a well-developed Longitudinal Data System. According to the Data Quality Campaign, states such as Delaware have all ten essential elements, which are the, according to the Data Quality Campaign, the ten

essential elements of a fully functioning Statewide Longitudinal Data System, and so Delaware has done that without any grantee funds.

It's important to note that although we do fund a lot of the infrastructure for this, at least for the start-up, in some states, you know, larger states, we, the price tag for these systems, at least the start-up, is not covered by the grant funds, and we don't want to be the sole provider of the funds for these systems. There needs to be a state buy-in for sustainability. So, especially in California and Texas, we are providing funds for maybe a portion or a certain activity.

Even in, quote, "smaller states," I mean the price tag for these systems can be pretty expensive. For example, Maryland, for their P-20 system, from early childhood all the way up to postsecondary, they estimated it would cost about $20 million. And obviously, California, just their K-12 system is much larger, between 30 and 60 million, just for the K-12 portion.

So that's the infrastructure start-up costs. The maintenance costs are still varied because these systems are, in some cases, a couple years old, so the maintenance costs have not really—there's never been a cost analysis about what these systems cost to maintain.

Some of the additionally collected items beside the student demographics, kind of the typical data that you would collect at the individual student level. At the state level, as you can see up here, several states are starting to put in the National College Readiness Assessments, ACT, SAT scores. Classroom grades are getting imported into these systems. Enrollment. Attendance on a daily basis, which is impressive in terms of an

organization from having a school input the data and then having that exported all the way up to the state level on a daily transactional basis.

Discipline. Enrollment. Homeless. Migrant. So you can see a lot of different data that they're collecting at the state level. And to give you a picture of how all this works, most of the districts, large district systems are more—a lot of the large district systems are much more developed than the state system. So the districts have been doing this for years, as many researchers realize. The state systems are trying to play catch up in a lot of these states.

So this data might be collected and lots more data might be collected at the local level, but at the state level, this is a growing trend, to try to bring this data up at the state level.

Then, in terms of our grantee states, and this is just the 42 grantees that we look at, I wanted to talk about some of the access to the student-level data that is given out. So including district staff, the parents, what grantee states have publicly published policy on data use. So I want to just give you some of the numbers from our grantee states that we recently asked.

So, for example, the access for the district staff at the student-level data, for at the state level, right now that number is 14 grantee states have—and it's operational is the top bar—14 states have an operational access for district level staff. 13, I think it's actually 18 are in progress to do that. Nine have not begun, and one have no plans for giving access to district staff.

And district staff—that is a large group. I mean you could consider teachers in that. You could consider district administrator of curriculum. You could consider superintendents, but still you can get an idea for who's getting access to this at the district level.

Access for parents. What type of parents have access to the student-level data? Right now based on our grantee states, four states said that they have an operational system that gives access to their parents for their student-level data.

Six are in progress. We've got 18 that have not begun and 14 are not planned. So the access to parents is an interesting question, and again this is self-reported by the grantee states, so what that access involves or what data they get to see for their students does definitely vary, but this is just an overall picture for who's getting access.

In terms of researchers—I probably should have led with this slide—but the publicly-accessible policy. So do they have a publicly-accessible policy on their website? Right now we've got ten states that have an operational publicly-accessible policy. Seventeen report that they have it in progress; they're working on it. Thirteen have not begun, and one not planned.

And when I get to the anecdotes about the four states, you'll see possibly why the district staff, there might be more access to district staff than researchers. But the 17 in progress is probably an important number to look at because they realize the interest of the research community and how to reach out to give access.

This publicly-accessible school grade level achievement. It's growth model data. Who has access to this based on individual student growth measures? Fifteen said that they have it operational. Eight in progress. Twelve not begun, and seven not planned.

And when I look at this slide and think at the state level, having access to growth model data, I think I would just caution that this is self-reported data, and this is many times, when we gave out this questionnaire to the project directors, there is a question about what this means or what is the publicly accessible. So I would caution that 15 states that might not have a readily accessible data file, that they would be willing to transfer with an MOU.

So this is the downside of self-reported data. So we are working on making sure that we explain what this is so it gives it a little better results on who has access.

Examples of an SLDS sharing with research community. So I picked these four states, and we do notice that Florida is not there. So I want to decide—I want to talk about some of the states that might not be always at the forefront of whether it's in the CALDER group or whether it's in the—I guess the usual suspects for giving access to student level data.

In Arkansas, they are very proactive in trying to reach out with the research community. They have MOUs with five agencies, research communities that they regularly give access to data. It's an ongoing relationship.

They have two websites that they give aggregate level data which

is mostly the requests that come in. And they have an interesting relationship where the project director actually on a request basis. He has this data file going back to 2004 with grades, discipline, demographic data that he has created, de-identified, and he will give this out to researchers provided there's an MOU, and he's a researcher himself. He realizes the importance of getting this data out to researchers.

So it's really, it's almost like an ad hoc. I mean how does this work? I mean who's asking for data. That's how they develop their system.

In terms of internal resources, they don't have several staff that handle the request, but they do have, they try to make this data accessible to the different groups.

Utah was probably the most surprising when I started to get down to what do they give access or how do they give access to their data? Utah has all ten essential elements. They're widely known as having a pretty advanced Statewide Longitudinal Data System. So in talking to the project director, you know, I thought there would be several advanced policies, procedures, possibly websites, and they do have a website that does give aggregate level data, which most states that have advanced systems do.

But I said do you have like an online application that researchers can go and apply and ask for data or do you have MOUs with universities? And he said, well, no, actually we don't have, we're starting to get an MOU with our local university, but it's been about 12 months in the making, and we don't have any online applications. We're thinking about trying to do that, he said, but honestly we don't have any requests for data.

You know it's from the research community. I mean they know about it; it's just that there's not a strong relationship, and so it's not something that they're holding back from this conversation, of course, but it's just a matter of they just don't have the need so they're not going to set up their system for the research community.

Moving to Rhode Island, going—an opposite approach—they actually have several access, several policies in place, several MOUs with local universities. They have an online training for researchers that can attend for accessing their system and understanding the individual student unit record data.

So it's a much interesting infrastructure of how they do it. So then I asked, okay, so the online training, this day-long online training: who comes to it? Is it mostly researchers? He said actually that's a small minority. It's mostly principals and business managers that come to this online all-day training. I'm sorry—all-day training, not online. It's all-day training to figure out their system. He said it's a very complicated system. And so it was interesting who comes to the training.

Louisiana. Probably one of the top three in terms of research requests. Again, they have a much more advanced organization. They have a hotline that people can call in terms of getting data, in terms of asking questions. They have six people that staff this office so when people ask, you know, how do I get access to data, the hotline—the six members are all—they're trained in terms of where to feed them. If it's just an aggregate level report, then you can go to this website to pull this down.

If it's individual unit record data, they will feed them to the SLDS team, and then I asked who are the six, the six members that are feeding these requests? And they said they're all university students that are social workers, teachers, researchers, so they—so it's a really interesting front-line experience for these six people that are working this office.

And again they have websites. You can pull aggregate level data. So it's a really interesting, just out of the four states we have listed there, it's an interesting—really it sounded like, you know, where is the demand? If the demand is there for researchers, we'll put an online application. We'll put a system in place, but most of their requests are coming from evaluation purposes, from other agencies in the state; principals; district staff.

So it was an interesting way to look at how they're doing this. Now, Florida, as we know, Florida—as you may know, they have a much more advanced system. They have staff in place to handle the many research requests. They have partnerships with universities. They're a member of CALDER with Jane. So they've been doing this for several years, and so, but again talking with their project directors, it is, you know, it's a cost issue. Their main portal they just built were for policy legislators.

They just backed off a parent portal because the funding is tight and they have to address the audience that is knocking the loudest in a way. So those were four anecdotal evidence of research.

So we've got—I also listed some of the issues identified through the program FERPA. I don't know if I need to elaborate on that, but the state procurement issues. Just in terms of building these systems or in terms of

building, let's say, a portal for researchers. Although we give the grant funds to the states, the states set up their own procurement processes for their contractors so they have—sometimes that can take up to 12 months to get a contractor on board.

So even though we just announced the 27 grantees, next week there's not going to be a Longitudinal Data System. I mean in some states, in California, it took 18 months to bring on their vendor. So but now they're doing—they're definitely moving along.

In terms of the third issue that's identified is the SEA as the facilitating organization. In terms of my phone ring, I think this is the one that has been the most popular is, 'are SEAs the only ones that can apply for these funds?' What if a research organization or what if a postsecondary agency that's much more advanced in these systems or what if another agency in the state would like to apply for the funds?

And with the recent focus of the stimulus funds of building P-20 systems, the SEA is put in a position to try to manage not just the finances but breaking down these silos. And that's a challenge for the SEA as the facilitating organization.

I guess I should answer that now just in case that question comes up later. Yes, SEAs are the only applicants, but they are the fiscal agent. But we do encourage them to work with other agencies, organizations, and on our website, which I'll show you in a second, there are partnerships with existing research organizations.

So the SEA will apply, but there will be a certain amount of funds

for a research organization to work on, let's say a report or designing a portal. So there's definitely a partnership, so it's not just we write a check to the SEA.

So sustainability at the state level. I mean I could elaborate on that, but I think we all realize, especially in these times, how difficult some of these systems could be to maintain, but I think it's just more, and the project directors realize it's about educating the purposes and the benefits of the system.

I mean this is not just for researchers. This could be for parents. This could be for students. I mean I know there's a superintendent in a district in Oklahoma, and the first thing he does every morning is he opens up his portal and he has all of his Longitudinal Data System, his dashboard he looks at, to see how students are doing, attendance, any drops in discipline. I mean it's a data decision that he makes every single day based on the Longitudinal Data System.

So it's how do we integrate these data systems in education, which is typically a paper-based profession. So as a teacher I remember very few decisions were made, even though I was a math teacher, about how to use data. Just it wasn't there. It wasn't in the capacity of how teachers would use this. It's not readily accessible right now, but it's a lot of districts are moving into very creative ways of getting teachers involved, and so the states will follow.

So the next steps—and this is the next steps that we see in terms of what the grantees are putting forth. Getting, creating research across—I'm

sorry—creating access across state lines. Regional collaborations. Postsecondary and labor linkages, and we should add early childhood especially with the focus on the stimulus.

And then providing data access to research community and public stakeholders. As I mentioned in the anecdotal evidence, a lot of these project directors don't realize how to give access to researchers. Arkansas was unique. One of their project directors is a researcher, but in some of these states, they just don't know what, yeah, give access to researchers, what does that mean? Does that mean we have to set up an MOU?

It's a lot of questions. So it's, they realize it's important. They just don't know how to do it in some cases.

Briefly, I'll go through the appendix slides. This is our website that we have that has links to our standards and guidelines, events, and presentations. We have a list of our grantee states. So you can click on any of these states and you can see how much grantee funds they've received. You can also click and pull down their original application so you can see exactly what they propose, their time lines, their budgets, who's working on the project, and we also have their outcomes because these applications are sometimes unwieldy.

So we put right on there just bullet points; their five to seven outcomes; their main focus for their grants.

This is an example web page of a state. This is Kansas. So it's probably tough to see. Let's see if I've got—this would be their major outcomes in this section. Their application you can pull down in PDF.

They've got their grant funds. We also put their project director—Kathy Gosa. We did not put their contact information for obvious purposes, but this is—and every state has this web page—and they also have their, in this section, let's see, at the top, any relevant websites they want to post let's say to their access to their aggregate-level web page or their Longitudinal Data Systems. So they provide us links.

This is actually an outdated features matrix, but in the questions that I discussed earlier, who's giving access to researchers, who's giving access to parents, we actually have all this on our website; and we have the columns [which] are the grantee states, and then down the left side, we have all the different features that we ask including some of the features. Who's collecting homeless student data? Who's collecting migrant data at the state level?

And the charts that I provided earlier that said "operational" and "progress not begun," that's where I pulled all this from. So this is a public document that we have on our website.

And I think questions we'll take afterwards; correct? Okay. Thank you.

MS. HOFFMAN: Thank you, Tate.

Sean Mulvenon.

MR. MULVENON: All right. Good morning. Thanks for coming out early this morning. That was some nice weather at about 6:00 a.m. Wasn't it? Did everybody get up early?

[Laughter.]

MR. MULVENON: A little background on me. I know that Lee went through some of this. I am a Professor of Educational Statistics at the University of Arkansas, and for a 31-month period, I was on an IPA at the U.S. Department of Education working on longitudinal data models. I was on the growth panels and got an opportunity to really kind of study what was going on around the country in terms of these types of Longitudinal Data Systems and what people were generating, and hopefully I can share some of that expertise or some of the knowledge I gleaned from doing that with you this morning.

Just for clarification, I am not part of the SLDS in Arkansas. I am a professor at the U of A. We do have a research center there that does a pretty comprehensive job. We do all the NCLB calculations for the state reporting, EDEN. We have data systems that distribute. We have a private and public side. The private side, teachers, parents, not parents, but teachers, anybody in the educational system can go in and actually track down their kids, and we've got longitudinal data that goes back to 1996 on these kids.

In fact, we just used that data recently on an engineering study of successful engineers through their second year. We tracked them all the way back to fourth grade and found out where they were sitting, who was on either side of them, and now we're tracking those kids forward to find out what happened to that pool of kids and why they're not in engineering. So, but any way, a lot of great stuff out there. We want more engineers.

All right. I heard this said at a talk I gave one day. We need to use value-added analysis. And my immediate response was why? And then a

lot of internal questions and questions I asked, but basically I was trying to gather an understanding of why the perception they had to have growth models? Because education is just a veritable cornucopia of clichés; right?

We need "value added." It needs to be "student oriented." You know, all kinds of two-word catch phrases. But the most important question, and I think it's a research question, and it's appropriate for this conference, and that is what are you trying to do? Because that really solidifies where you're going, what kind of data you need, the depth and clarity of the data.

Now, when I review a lot of what's going on with the Longitudinal Data Systems and particularly the use—and a lot of this is just off some of the websites and some of their sites for articles that have been used or generated using their data, and there are some great things that are going on out there, and a lot of great ideas, and it's very encouraging. Always the "but," the other shoe—right—but there are some really interesting problematic things that are starting to occur.

And I think that a longitudinal data set, I mean it's a great thing to have, but it has a utility, and it has limitations, and I'm going to talk about some of those. There's some incongruence in the reporting models, and I think this is really troubling and creates a lot of questions when people are looking at cross-sectional data that may be reported as part of NCLB versus longitudinal data that's also reported as part of NCLB, and they can't make the connection between the two.

A good example. In some of my advanced classes, one of the assignments is the students have to go and get an article where they have a

correlation matrix and means and standard deviations in the article because if they provide that, see, we own them. We can replicate the models that they ran in their study and see how they reported it.

And I hate to break it to you, but about 50 percent of what we find is it's not: either hasn't been analyzed correctly or it isn't reported correctly.

But we're replicating what they did. So we know there's a problem with how—and these are people who are supposed to know what they're doing. So I can only imagine what's going on in the educational, at the school level, where they just don't have that depth of understanding.

Back to what are you trying to do? Obviously identify research questions. Develop appropriate data sets. Select the appropriate analyses—appropriate analyses, not all analyses. There isn't a one-size-fits-all in analysis.

Oh, and a heads-up, I got way too many slides. So I may start going fast. So I apologize for that.

Anyway, I see it's the fifth slide before I even get to the goals. Goals of the presentation. Okay. What are LDSs' implications, strengths, weaknesses, limitations, challenges to developing these data sets? And how they can be used to expand research capacity in the school systems is what I'm hopefully going to try to do with the rest of my 700 slides.

Now, issues that must be addressed? Matching. One, when we were on the Growth Review Panels there at the department, we always got these astronomical figures of these match rates that people had, and there was

just no way, just the volatility, and those were red flags. And so what do you mean by matching?

Merging is an interesting question, too. People think you have data set one, data set two, boom, I got a longitudinal data set because this is year 1, this is year 2, when actually when you create a longitudinal data set, if you really break it down so it has a functionality that is both horizon and vertical, and I'll talk about that, to add to this congruence, it's actually a composite of seven data sets that are concatenated at the end. And I'll talk a little more about that.

Functionality and then data quality and what we mean by that. What are you merging? Okay. You know, it's funny, you ask people like a unique student ID. I heard that mentioned yesterday. And if we could give them at birth, and all I could think about when I heard that, 'if we could give them at birth,' is when my daughter was born before I could leave the hospital, I had to fill out the Social Security card thing. So we do give them a number at birth.

So we got them in the system. We just don't have any data on them, but there is some different examples, probabilistic neural net. That's what they claim they use in Arkansas. Interesting side point, that we have higher match rates the way we do it at the U of A for the longitudinal scoring than they actually do in their LDS system with the probabilistic neural net. That's a different paper.

Bashing. You just slide them together. Merging on multiple variables. There's all these different things that people do, but how do they

work? How do you handle redundant values or duplicates? Should you take the first duplicate value or the second duplicate value or the third? Which one is the most appropriate?

There's a studying that has to take place. Here's an example of something that happened when I was at the department. I get summoned to the Secretary's conference room, and they're all distressed because one of the NCLB Growth Model states only has 500,000 or so kids in their growth model, and they can't understand this because there's 1.3 million kids in their system, and they got a 99 some odd percent rate.

Well, I walk in the room, and my first response was, wow, that's really good. What do you mean they only got 500,000 kids? I'm like, well, let's think about what they should get. And if you really think about it, they were using grades three through eight. Third graders move to fourth grade. Third graders don't have a growth model. Eighth graders exit the system. They got five grades, about 100,000 kids in the system; they say they got 495,000 kids. That's pretty good, and that's what they should be shooting for. So understanding the context of what you should get when you build your data systems.

The horizontal and vertical functionality. Here's a horizontally functional data set. Looks pretty tame. I can subtract. Those are reading scores in third grade, math and reading, if they're scaled properly. Operative word "scaled properly."

Then I can subtract those two and kind of get a difference about what's going on versus reading and math. Unfortunately, this is a fantasyland

model. It doesn't work that way because of the distributional assumptions on the different tests. So it can't be that clean.

All right. Vertically functional. Same type of thing. I can sum up and down as long as there are some certain scaling features that are congruent. And when I looked at what was going on and what's being encouraged, I'm not always finding that people are being educated on some of these underlying metrics that you have to adhere to if you're really going to get meaningful information out of your LDS.

All right. Now, it seems obvious—right—subtract column one from column two. Got it. We can all do that in Excel. But here's a MYSQL data set. All right. Typically, when data sets are provided by these, you know, too many of these—if I could lobby for anything, I wish the LDSs couldn't use external contractors but had to build it internally, to build their infrastructure as opposed to just putting money into contractor pockets because—or at the universities where they could build it because you have that long-term sustainability.

But when you get your data sets, like when I was at the department, we'd get data sets from the EDEN system on Perot that would look very similar to this data set right here.

Problem is, this isn't a functional data set for statistical modeling. You got to reprogram this. You got to write the code. I have several thousand lines of code to reformat the EDEN data so we could actually analyze it. But then when we got all done, it was pretty powerful.

By the way, the first people that asked for a copy of those

programs were Perot, but they're housed over at the U.S. Department there, but you build the data sets because they use these for reporting features. This is a matrix and they use Reading 37. They put that matrix cell, and that's how they put it on the reports. That's how the data sets are used.

Assessing data quality. This is a good one. I did this with Arkansas data. Third grade to fourth grade, I found the perfect 31,000, 30,000 cases, whatever it was. 30,000 cases. This is clean data. Everybody's got a free and reduced lunch value in there. And then I perfectly match those kids to fourth grade, 100 percent match. I mean this is just clean data. It's as clean as you can get it; 100 percent. So everybody excited. The data quality people are thrilled.

But if you cross tab this, looking at free and reduced lunch, 12 percent of the kids change their values. So you got a great data set. You've got perfect data in year 1, perfect data in year 2, and that variable is too volatile. It can't be right. That's about twice as much as we'd expect.

See, so something is going on. Oh, and by the way, why we picked that variable was because it's adults that are filling it out at the school systems. See, so it isn't some kid coded it wrong. This is just something that's structurally questionable so you got to go back and understand it.

Anyway, oh, and here's the great four dollar question. I saw in the poster presentations—by the way, Florida State, if anybody is in here, awesome job with those pre-doc students or the doctoral fellows on those papers. They were outstanding. But there was one in there where they were talking about the assignment status, and I just, I loved it because this is a real

question.

In your LDS, which assignment status do you use for your kid now for FRLP if it's changed? Do you use the 2006 or do you use the 2007? Real questions, and they have implications. Okay. I can tell you what the state had us do.

Here's an example of other things that we see, and it goes to that data congruency. See, this is more like what a real data set looks like. Would you guys agree? There's missing values, kid moved out of the district, came back in. All right.

Now, what typically happens is they report model one means, and that's basically they just sum the columns, and those are the values; but when I do a longitudinal model, I'm only going to use observations one, two, three, five and eight because they're the only ones with complete data. And so actually it's the means in model two that are the ones that are actually used to calculate the results.

We see this a lot in the misreporting of data in journal articles where they do the descriptives at the front end. Then they run their models and all those folks without complete data get kicked out, so actually their analyses are based on a whole different set of means and correlations that aren't reported in the actual analysis or in the article. And that's where a lot of the problems occur.

Meaningfulness. Getting people to understand this. Because you look at some of the distributions that are reported, the standard deviations, and this is a slide I use to help people understand that.

Now, reasonably educated people that we are, we understand that statistics isn't about just the difference between two means. It's the difference between two means predicated on the distributions of those different variables; right.

So at the top, that's an example of it's not meaningful—by the way, this slide works really well with educators to try to get them to understand this—where at the bottom, that five point difference is meaningful, but it isn't always a five point is meaningful, and it isn't always a five point isn't or is not meaningful. It's predicated on those relative distributions to that in particular study.

So, and an example, and I'll show you another example in Arkansas in just a second. But it's not just a list of variables. Like DQC does a great job of championing. We've got to get these ten variables in there, and I think Aimee Guidera is one of the best there is. She's head of the Data Quality Campaign and presenting a lot of this to the states, but I also think that having the ten variables doesn't necessarily make it great either.

It's what you do with those ten variables, what you get them, and how you control for the validity of the data sets.

Growth Models is a field in statistics. All of those are potentially appropriate methodologies. It depends on the data sets, what you're trying to do, what you're really analyzing, but it doesn't have to be an HLM. Most people aren't even really sure what an HLM is. They just know they need it.

Latent growth curves. Yeah, my wife is an elementary school principal. I've heard this for years; right. Latent growth curve modeling. I

love latent—I prefer latent growth curve modeling. I think it gives you a lot of different types of trends that you don't necessarily get in some of the HLM modeling.

Florida State had one of those papers, and it was outstandingly done so, good for them. Our new Ph.D. program aspires to be like the one at Florida State. Let's just say that. All right.

Two major models I looked at: equipercentile and growth trajectory models. Now, an equipercentile model is like North Carolina is using an equipercentile, and here's an example of how that works, and then I put a metric around it. This is a different thing that we used in other states that we worked with and districts.

My daughter's third-grade teacher told me you couldn't help—you couldn't improve the scores of the advanced students. It was kind of an interesting comment to make to me of all things, but this woman is, she's won a couple of math awards, and she is sitting across from me, and she's got two posters behind her, "Math Facts." The irony of the two math facts is they were both incorrect.

It's a true story. But she said she couldn't help the advanced kids. This is a misconception because you have to look at the distributions again. So, in reality, this is a great one for policy where you see a lot of policy misinterpreted.

See, the student at the top if kept a relative position would be expected to gain 20 points. Goes up 22. Student at the bottom, expected to gain 40 points, gains 36. So the relative view of this is that (a) they're

closing the achievement gap. That's not true. And that student on the bottom did better than the student at the top. Well, relative to the distributions, in fact, the student at the top did better than the student at the bottom, and the gap is getting wider.

But they're using the relative scores, not indexed against the distributions. And we're seeing a lot of this in the modeling that's going on. A great place for research where they do stuff like this—it's crazy—is AERA. Okay. Lot of stuff like this.

All right. Value added. It's a simple slide to show people what they mean by value added. In terms of predicting out, and you're looking at the residuals, those are projection methods.

Growth Model. Develop goals. I think we all know this. I know my time is running short. And I think these are all good things that we want to do. We want to identify student improvement. We want to predict performance, to get an idea about where the student is going to go. We want to evaluate curriculum and professional development. We want to develop things that can help us work with teachers.

Another important input in all of this is the validity of the exams. The psychometric properties of some of these exams can be very troubling, so no matter how sophisticated your analysis is on the back end, if your input is a little problematic, then you've got some issues; right.

We've got some standard errors that when I was at the department for states on their exams that are wider than their performance intervals. Okay. So you've got some real issues with how clear the exams are

psychometrically. Are they vertically equated? Arkansas, they use something called vertically articulated, and I'm going to show you what that distribution looks like.

For accountability, there's growth models and there's scoring models. And they should not be confused. Growth Model is a methodology we're using. Scoring Model is how we take the Growth Model results and convert it to a way to assign a school a category.

And the Growth Model, the base math is pretty solid. Scoring Model is where it gets a little creative. And you got to be real careful. For example, a declining score—and I'll show you an example of this here in a just a moment—a declining score where a kid is clearly not growing can be counted as meeting adequate growth as long as they stay above the proficient line.

So if you run an analysis and you're looking at kids, and they got declining scores, that's going to confound your results relative to the actual reporting in NCLB, and I'll show you an example of that.

Formative measures. People are using a lot of additional formative measures to drive this. I think that's outstanding because these large-scale standardized tests—wow, that's a lot more time than I thought—okay.

Anyway, these large-scale standardized tests, truth be told, are not designed to be diagnostic. In fact, they're under-parameterized for that purpose. All right. So the idea that if I look at the 40 multiple choice items on the—in the—eight or the five constructs on the math exam, in the state of

Arkansas, I can diagnose everything that's wrong with this child and why they're not learning math is a little untenable.

But the infusion of formative items, additional items at the classroom level linked into the longitudinal data sets, which, by the way, in my research office, the NORMES site, we seed all the data. Any of the standardized achievement test data, if you're a principal, you log in to the ED stat portal, and all your data sets are seeded on the side. Now, the nice thing about this is, is we kind of know how to build data sets so it's transparent; it's all there.

Anything we're going to do, it's there, you can request it, but more importantly because we use the Business Intelligence Suite through SAS, we got unlimited access to what's called "Enterprise Guides." And we engineer Enterprise Guide, and so the principal can log in there, go into e-guide, pulls in their data set, and then they can do any analysis they want, and it's point and click, drive. We got support teams that work with them on it.

But it's free. It doesn't cost them anything because it's part of our system at the U of A. All right. So it's a real nice way for them to do research on their own systems. Furthermore, it's a wonderful way if you're an educator and you want a professor from Arkansas State in Jonesboro to come into your school system, there's an analysis system. There's your data. You can drop and drag it right in there. You can look at it and it's as clean as you can make it, and it doesn't cost $20 million.

Okay. It's a much cleaner way, and it's a way that I think we can get more researchers to use the data systems because just the ease of access,

and the principal/building administrator can, of course, give them access or priority if they're so inclined.

Some interesting questions in regards to the methodology. Student matching. How are you going to do that? Covariance models. I saw a wonderful presentation poster where they were explaining all the ills of education. If you controlled for SES and race—okay—that's true. There are some clear correlated trends between SES and race in terms of achievement.

I'm always troubled by that, though, because I can't control SES and I can't control race. Right. I don't know what the intervention is that I do to control race and SES. I know they're different. What are the instructional strategies we can do to address that as opposed to just isolate it, as that's the excuse?

Because I can tell you right now, working with principals and teachers and things, as soon as they hear that, it's almost like they throw in the white flag and say, well, I can't do that. The kid is poor and he's black or that kid is poor and he's white. Well, whatever the issue is, and if so, you need to really think about the use of those demographic models.

We know there are differences in performance patterns. The challenge is what are you going to do instructionally to try to eliminate differences associated with those different demographics?

What are the decision rules? How are you going to score all this?

Here's some outcomes. Professional development and reporting results. I think that's a major goal of these people with these LDS systems is that they want to improve the reporting, they want to improve the data

availability, but more importantly, they want to link it to professional development that will really help improve school systems, which I think is awesome. That's what we really need to do.

Here's an example, and I was talking about the gain systems. Here's the Arkansas—if you're a third grader going to fourth grade, you're expected to gain 59 points in literacy and 59 points in math because they're on the same scale. Here's a problem. The standard deviation in literacy is about 90 points, and the standard deviation in mathematics is about 15 points.

So a kid goes up 30 points in literacy, and he goes up 15 in math, so they go, by God, you got to do better in mathematics when the math effect size is about half a standard deviation, whereas, in literacy it's a third of a standard deviation. They got the wrong target.

You guys follow? Now do this across time, and we're seeing too much of this, because the ability to point and click and drop and drag is great. And these data sets are awesome, but there are some underlying statistical models and tenets that are just kind of being dismissed, and that concerns me.

Look at how the requirements decline across time, too. Because it's autoregressive and nonlinear. That's the distribution across time. So if you don't address that in your models, and people don't, then you've got a problem. The interpretations aren't—we're not sure about how accurate any of the outcomes are going to be.

Just full of good news this morning, aren't I? All right. I'm meeting with the Arkansas Department of Education tomorrow to share some of this with them, too.

[Laughter.]

MR. MULVENON: I'll call and let you know how it went.

The transparency is also a big piece of this. Make it as transparent as possible. Right. So the reporting—these were some growth reports that we have, that we put out actually. Here's a student report level that a principal can get.

Now, I don't know if you can see it, but if on the first, second, third, fourth column it says expected growth.

If go down that list, you're going to see kids with negative growth expected values. Those are high-performing kids, and they use this trajectory, and the bottom line is the kid can decline. If he declines by 70 points, hey, you still made it. See, that's a problem.

Now, I take all that data and I analyze it, and I'm using this, I've got issues because it's a confounding result. All right. So it creates some interesting conundrums in thinking about how we're going to do this.

That's real, by the way. And, yeah, I lobbied long and hard at the Department of Ed when they were making the decision about whether to allow people with the declining slopes to be included as proficient.

I lost. But I fought the good fight, and it was a policy decision because they felt that they were moving it forward to get growth models introduced into the terminology and into the process, and I got to admit they were correct. It's been beneficial. But these are things I think in the second iteration that you address.

Subgroups and how they perform, I think that's always good. All

right. These are questions that people are doing, which I think is good. Okay. Thank you, Lee.

Interesting. Evaluating why students did, did not make expected progress. These are all outstanding types of things that people should be looking at, and they are looking at, which is good.

Which students at the classroom and student level? I also think you got to be real guarded about individual student evaluations. Of course, we want to look at how individual students are doing. But the unit of measure really is the classroom level as opposed to the individual student level.

So be cautious about that, and I get real sensitive to that because you hear these stories about, well, you know, Johnny is measured in six categories. He's poor, he's white, he's, well, he's free and reduced lunch. He's in a special ed program so—and he's in the combined category.

So for math and reading, Johnny is measured eight times, but NCLB is actually about groups. It's not about individual students. It's actually about groups of students and whether they're underserved. So it's an important thing to examine.

I think it's—here's some quick examples I'm going to show you. Professional development; public reporting of school performance—that I think are really, they're good. They're good examples of where they're doing some things with these LDS systems to look at how we're doing in the schools, and there are different ways you can index this.

Now, here's an ITBS literacy exam. Now, this particular school district got a grant from a company called Wal-Mart. Anybody heard it, heard

about it? Pretty much everywhere now.

Anyway, but when they were asked about the methodology, it was put on that they had to do fall/spring ITBS, complete battery. So they had a baseline when the kids came in and an exit point when they level.

Oh, and we didn't do the testing three months before the end of the year. We did it 2 weeks before the end of the school year and 2 weeks after the school year began. So we got a really good blocking of what's going on, but look at this trend here; it looks like it's a long tail to the right.

So they're comparing these kids. Anything above 70 is good. Because you index against something. You got a—a key of all of this is we got to convert it from the HLM and the equipercentile models and all this language that the layman doesn't understand into things that they do understand.

And of the things that we've done is we've taken this data, and we've renormed it and transformed it into these values where 70 is the norm. We picked 70 because everybody knows 70 is passing so if you're above 70, that's good. If you're below 70, that's bad. Wow. I can walk in a room in a minute and explain that to parents, and that's all they need to know.

Okay. So in this particular case, they're well above 70 when indexed against the national norms. But see them declining off. Patterns and trends are important in this game. It's the same in mathematics. So it wasn't just isolated with literacy.

Something is going on in this district. But what a great target point for their curriculum people to get involved and start figuring out what

this is. The tendency would have been just to go down here to tenth grade and go, look, we're above the national average and quit. But look at that steady decline, regardless of whether it's literacy or mathematics.

Here's an example—real case—we talk about at the teacher level, I know that the—okay. Thank you. I'll wrap it up here. Just one more, two more slides. This teacher—this is a real case. Now if you look at those reading scores down there, and she's doing a horrible job. So what's—oh, she's that bottom ten percent they were talking about; Secretary Duncan was talking about yesterday.

Well, actuality, that's year 1, year 2, and year 3 of that teacher's teaching career. She was a newbie. The teachers two and three had been around for a couple hundred years. They were pros.

[Laughter.]

MR. MULVENON: Okay. The principal was able to work with one of those teachers to help this lady, and they sent her to reading programs, and she transformed her reading. She was an excellent teacher; she was just new, needed some help.

So these are things that if you are proactive with these types of models, people are much more excited about using and linking it to teachers because it's not used as a hammer; it's used in a productive way.

Example of a school district that isn't doing as well. A school, well, they're above the national average, but you look at their growth relative to one that's doing really well, and they both have the same growth. Not closing the achievement gap, but doing very well. So, again, indexing and

making it relative.

There's, I think, growth models definitely work in education. There's a lot of exciting things that people are trying to do, and the one caveat emptor thing I wanted to share was we got to do a better job of educating people on some of the methodology concerns and the statistical parameters that go behind it, beyond just building the data sets.

Thank you.

[Applause.]

DR. HANNAWAY: Hi. Thanks for inviting me.

This is, I think what I'm supposed to do here is talk about what sort of findings are coming out of these State Longitudinal Data System based research that have implications for policy. And that's what I'm going to do.

So, first of all, let me just set the stage a little bit. I think everyone agrees the U.S. education system is in trouble at the system level. About half of minority kids, only about half are graduating from high school; four-year grade gap between white and minority students by 12th grade; 15 percent of minorities getting BAs 9 years after ninth grade.

Everyone agrees: trouble. In order to get anything done, we need both the will and the way, and I'm not trying to make this into a religious presentation, but I think the will is there—my own judgment. It seems that there is pretty much agreement on the left, the right, and center that we do have real serious problems in education; lots of strange bedfellows on trying to figure out solutions.

And the problem I see is the way. So I think the political will is

there if we can show the way. And we have few guideposts about what to do to solve the education problem, but we do have a few.

And the way we're finding the way is through these State Longitudinal Data Systems. Obviously, they can't answer all questions, but I think they are giving us some important insights. So states have the makings of the evidence in order to indicate how to proceed, and I think it's one of the most important effects of NCLB, is the creation of accountability systems and the creation of these Longitudinal Data Systems.

What do we know? This is what we know so far, and these are findings that have come out of research using State Longitudinal Data Systems. I might add that these findings are CALDER findings, as well as findings by other researchers. They're findings that hold across states; they're findings that hold across tests; they're findings that hold across multiple studies.

So they're pretty solid findings, and they're giving us the basis for beginning. First of all, teachers matter—single most important schooling contributor to student outcomes.

I remember the day when around schools of education, there was a lot of talk—and in the social science research community—that schools really could not have much impact on student outcomes; that student outcomes were driven by student background characteristics. And certainly student background characteristics are and what goes on at home is an important piece of the puzzle, but it's not a determining factor.

Indeed, there is wide variation across teachers in terms of their

effectiveness and in terms of the gains that they get from students, and this is really, really important because teachers are getting these gains with poor kids; they're getting these gains with minority kids; they're getting these gains with white kids; they're getting these gains with middle-class kids.

It's the teachers that are driving it. Some teachers are just a whole, whole lot better than other teachers.

And this gives us a key insight into what we should be doing in terms of education policy because education is fundamentally a human capital enterprise. It's labor intensive; it's people that drive the work. So those differences among people, both in who they are and what they do, are important.

We know a lot more from these data systems about who they are; we don't yet know a whole lot about what they do, although we're trying to team together with people that are paying attention to what is actually going on in the classroom behaviorally, as well as to who's doing it.

Now, the other thing we've learned, and this is something that holds across studies, across states, across tests, is that our standard measures, the measures that we're using to manage the human capital in this enterprise, the indicators of quality are weakly related, at best, to student outcomes.

So these are things like certification, years of experience, graduate degrees. They cost us a whole lot of money, but we aren't getting much payoff in terms of student outcomes.

So these—I don't think anyone would disagree with these three basic points, and I think they give us this set-up for where research should be

going if we really want to leverage greater student achievement, which I think we do.

So we know that teachers matter. We don't know a lot about what it is about teachers that matter. So it's our big, big puzzle.

I'm going to talk about three research probes. The research we do does not result in policy prescriptions, to a large extent, but they do result in better definition of what the issues are and, therefore, give us some direction in terms of how to shape policies.

And I'm going to talk about three areas of research coming out of CALDER: teacher maldistribution; teacher selection; and teacher mobility; and these are related.

One study that we're in the middle of right now, and I probably shouldn't even be talking about it, but I'm going to talk about it anyway, and some of these findings are sort of hot off the press, but I think they give an example of what we're starting to find.

This was a study that we're in the middle of for the Department of Education, where they asked us to compare the value-added of teachers in high- and low-poverty schools. We used both North Carolina and Florida data in doing this so we have very, very large data sets. We're doing it in two different states, again different sets of tests, different policy environments.

The basic findings that we have is that, lo and behold, there is on average higher value-added among teachers who are in low-poverty schools than teachers in high-poverty schools, but the differences are very, very small and often aren't significant in some of the comparisons.

I think this is pretty important because a lot of people would have predicted based on levels that the differences would indeed be very large. However, data also showed us that the variation within the high-poverty schools was significantly larger. The variation in teacher effectiveness in the high-poverty schools was significantly larger than it was in the low-poverty schools. This is important, I think.

Let me show it to you graphically. This is North Carolina elementary math, and you can see that when you look at the value-added of teachers who are in the 90th percentile, they're very similar in the high-poverty and lower-poverty schools. This isn't even looking at the extremes of the low-poverty schools.

It's looking at comparing schools that were at least 70 percent free and reduced price lunch to schools that were beneath that. And what you see is this spread is at the bottom, that the teachers in the bottom decile, those who are in high-poverty schools are a lot worse than teachers who are in the bottom decile in the high-poverty schools.

Now, this is important because if you think of a kid going through a school, a kid goes from teacher to teacher. Well, that poor kid. I mean that poor kid in the high-poverty school could have the knock-your-shoes-off teacher one year, and the next year have one of these really bad teachers and just really get pulled down.

So the kid goes year to year, and the variation within the school is important, and we know from most of our research that the variation of teacher effectiveness or teacher value-added, as measured by test scores—

yes—is about as great within school as it is between school. Important policy observation.

Florida. Look at this. Similar findings except that we even see that the top teachers in the high-poverty schools in Florida exceed the top teachers on average of the lower-poverty schools in Florida. We see it switches again at the bottom.

So there's something about variation within school that's important; there's something about variation within school that seems to be somewhat systematic. And this has policy implications.

You know, one, is in terms of the learning trajectory of kids as they go from teacher to teacher in terms of practice.

A second one is how we think about accountability policy. So if we have accountability policy that is confined to school-level measures, then we are slamming—some schools that we're saying are not making adequate progress or that are low-performing, there are teachers in those schools who are being labeled as being in the low-performing schools who are terrific. They're just as terrific as teachers elsewhere.

So it's really important about how we think of accountability, how we give credit where it's due, how we hold individuals accountable for what they're doing, and how they're contributing to the whole school.

General finding. Novice teachers are less effective than experienced teachers, and returns to experience tend to taper off three to 5 years out. Again, this is a general finding. It holds across states, across studies. It's here.

What is often not stated, however, and when you talk, sometimes when you talk to people in the policy community, they jump on this, and they say, okay, that gives us our answer. Let's go in, we'll have no more low-experience teachers in the most challenging schools.

Well, there's some sense to that, but here is the distribution. Again, this is North Carolina. This is in the high-poverty schools. The solid line is all novice teachers. The dash line are teachers with 1 to 2 years experience, and the dotted line is teachers with 3 to 5 years experience.

And this is the finding that we generally have that argues that inexperienced teachers are less effective than experienced teachers, but what I want you to pay attention here is not the movement of this distribution, which is moving up the productivity line there, but I want you to look at the overlap. The overlap is huge. There's lots of new teachers that are really quite terrific, and there's lots of very experienced teachers that really aren't.

So while we have this, you know, average movement, we don't have the solution. We're nowhere near the solution in terms of trying to understand teacher productivity and trying to develop policies that lead to more highly productive teachers in the classroom, especially in classrooms in highly-challenged schools.

This is in a low-poverty school. It's the same pattern of movement up. Looks like the learning happens a little bit faster there, but it's another story.

Maldistribution. I'm going to show some figures. This is work that was done by our CALDER New York partners, Susanna Loeb and Don

Boyd and Jim Wyckoff and Ham Lankford, in New York City, and there was a policy shift in New York City which phased out emergency certification. So emergency certification in many large cities were teachers that were—I don't know—close to dragged in off the street at the last minute and put in the classroom because typically the whole hiring cycle is very late in big cities.

So they said no more. And they opened up pathways for alternative route teachers. These are mainly New Teacher Project teachers and TFA teachers.

This is—and the characteristics of teachers in those schools shifted quickly and shifted fairly dramatically, and these are some indicators. So this is the failure rate of elementary schools' teachers by the poverty quartile of the school in which they teach. And what I want you to pay attention to is the blue line, which is the poorest schools, and you see the shift in policy. Boom. Big shift in who's in those schools.

Look at the green line. The green line is the low-poverty schools. Policy didn't have much of an effect there. These are the new teachers so when you look at who the new teachers are, coming in, you really saw a big shift in terms of who are in those schools as a consequence of a relatively simple policy change, in terms of selection and recruitment.

With what effect? Okay. And this happened fast. I mean we've been waiting decades and decades for effect. This happened fast. First two bars are the most affluent decile schools, 2001, 2005. Second two bars are the poorest decile schools, 2001, 2005.

What we see is—yeah—I mean the most affluent schools are more

effective, getting bigger student achievement, than the lower decile schools, but look at the shift. I mean the big increase in the poorest decile schools was quite large, and the last two bars show the gap, and what that indicates is that there was almost a 25 percent reduction in the achievement gap between those two sets of schools as a consequence of recruitment and selection of the individuals who were going into those schools.

I won't get into a lot about what they did, but you can see the characteristics of teachers who are in these schools and how they shifted in terms of, you know, passing the exam, certified or not, math SAT, competitiveness of college, but their characteristics alone shifted.

Again, the point at the bottom is really important. These are based simply on the attributes of those teachers. Clearly, any comprehensive program to improve student achievement, reduce achievement gap, would also include monitoring, development and selective retention.

Okay. A little bit more about teacher selection, and this is a study that we did relatively recently on Teach for America. Let me give you a word of caution here. When we started this study, we were working with the data from a large city district that had a lot of Teach for America teachers, really wanted this study done, and it was a terrific school district.

They were going to give us open access to the data, really wanted it, so we would have results, could turn it right over to decision makers because they want to know are we really getting much payoff here?

So we started working with their data. It took us months and months and months to get the data from them, and they put it together, and

it's difficult to put these data systems together, as has been suggested, because you have the HR files with all the teacher data in them, and you have the student files, and these are completely different silos or have been.

And you have to merge them and then you have to, you know, link them to schools, link them over time, so there's a lot of data manipulation there.

So we get the data, something just made me uneasy. It just made me uneasy. And so I went to TFA, talked to the district and said, you know, I want to confirm that the flags we have for TFA people in your file conform to who TFA thinks they have in the district. Because this is, I mean this is our critical indicator variable. 25 percent overlap.

So if we hadn't done that check, we would have come out with a study on TFA where only 25 percent of the teachers were actually TFA teachers. And so we talked to the school district, and they were, I mean this school district was terrific, and they were, you know, I think if we went, and they were moving fast, and I think if we went back to them this year, they'd probably be all clean and ready to go with this stuff.

But we went back to the school district, and I said, um, you know, I just got to jump, I just can't, you know, I don't trust the data, I just can't do anything.

So we jumped to North Carolina, whose data we had worked with a lot, and we knew the data well, and we trusted the data, where they have end-of-course exams, and so we focused on secondary schools, mainly math and science teaching, and estimated the effect of TFA secondary school

teachers in—relative to other teachers in North Carolina.

What I want you to compare is the top line is the TFA effect; the next three lines are teachers with more experience. The first column is TFA versus all other teachers. Second column, we're comparing TFA with other teachers who are also completely fully licensed in field. And as you see, in the first column, we get more than a TFA impact, more than twice the impact of experienced teachers.

Next one, nearly twice the impact of experienced teachers also licensed in field. We did a first version of this and got a call from the AFT. And the AFT asked us a very reasonable question, and one of the good things and bad things about when you're working with these data files is you can never give the response, oh, well, that wasn't the question we were asking or we didn't collect data to address that question. You're never done.

So AFT asked us a very reasonable question. They said would you get the same effects if you confined the comparison to teachers who had gone through a North Carolina fully-certified program? In other words, not teachers who were certified because they got a course here and a course there, came in from out of state, and, you know, basically got a mishmash. Very good question.

So we went and we redid the analysis because you know we got the data, we can do it. And we got the same effects, nearly three times the effect. So this is selection. There's something that goes on in selection that can have a big effect. Now, I want to—this study has gotten a lot of attention. I want to qualify it a little bit.

All we're talking about here is secondary school teachers, and mainly in math and science, and it conforms with earlier studies that showed that teachers who have stronger backgrounds in mathematics who are teaching secondary school mathematics get better gains from their students. So this is not sort of a wild finding, but what it means is the recruitment and—okay—recruitment and selection and assignment procedures can have a big effect on student achievement.

I think from a policy point of view, they also call into question whether we should be thinking about secondary school teachers differently from the way we think about elementary school teachers? Are there different characteristics that are important for secondary school teachers?

How important is experience? You know what is the balance between experience and expertise and how do you think about this? So what it does, again, it's a research probe. It opens questions as opposed to coming up with solutions. The solution is not have TFA take over the world. It's not going to happen. But there are, I think, some policy issues here that get raised by the findings.

Teacher mobility. Again, finding we have across a number of studies, number of researchers, number of states, and mobility is highest at the most challenging schools. That's where you see the most churning.

What we have recently found—and this is with data from North Carolina, Texas, New York City—the first to leave are the worst teachers. A lot of people always talk, well, the first to leave are those who have options. You know the most talented are going to leave. Uh-uh, it's the worst teachers

who are leaving first. This is a good thing. It means there is some natural purging that goes on, that teachers are making selections.

And it doesn't mean that these worst teachers weren't working hard. I mean teaching is—I did it for a year—hardest job I've ever had. It's hard work. And it's no fun to walk into a job everyday where you're basically knocking your head against the wall. So it shouldn't be surprising that there is some self-selection that goes on that where the bottom teachers leave. Not enough of them, I might add, but they're the first to leave.

There's also a general tendency across all teachers to, over time, leave the most challenging schools and move to the more affluent schools. Now, my point here on teacher mobility is that we can't get a good picture of policy prescriptions by just taking snapshots of schools because teachers are mobile. This is a dynamic situation.

So you can't solve the problem by getting the best teachers into the worst school and think 'done,' because teachers are going to move around.

Okay. The topic of the day. This is what everybody is talking about; right? Performance incentives. Well, the first question I would ask from, you know, the research we've been doing is, well, what's the objective here? What's the objective of performance incentives in education? What are we trying to achieve?

Are we trying to do it in order to improve recruitment and selection? I think this is part of the strategy of Michelle Rhee. She thinks if there are more rewards for productivity, that it will attract in a different teaching pool. I don't know if this is true or not. I don't think anyone does,

but it's certainly a theory.

You certainly change the risk profile of this job so there are some people who are not going to want to come into a teaching situation under those conditions. Is this good or is it bad? We don't know, but it definitely is going to affect recruitment and selection into the profession and we should be there to figure out ways it does.

Retention/deselection. Rick Hanushek and other people, like Secretary Duncan, are talking a lot about retention and deselection. So again these performance measures being used to reward and keep the best teachers in these challenging schools, for example, are—five—got it—are they used to—will they be used to weed out the bottom?

Third, is the purpose of performance incentives to induce greater effort? Teachers get insulted, and I think many of them for good reason, when you say the reason we're putting in performance incentives is because we want you to work harder, and they go what? You know many of them are working—many of them may not be able to do it. I mean it's really hard to do. We don't know what the magic is that these tremendous teachers have, but a lot of them are there, you know, knocking themselves out trying to do it.

So we have to think about whether it's inducing effort that we're trying to get at because it affects what we look for in terms of outcomes when these performance incentives get put into place.

Okay. There are issues, right, with performance incentives. First, how good are the measures? This is critical. There's measurement error. There is variability from year to year in these measures, all of which affect

how they can be used fairly and improve the system.

Issues about whether we should have individual or school rewards. The chart I put up before that looked at the variation within school makes us ask this question. What is the balance? Certainly there are costs of having only school rewards. We don't want teachers competing with each other. We don't want to reduce cooperation and learning across classrooms, so the balance between these two is critical. And what about the teachers without test scores? Lots of teachers don't have test scores.

Value-added measures. Again, we have problems of year-to-year variability in these measures. We have measurement error associated with these measures. A lot of debate about the extent to which the statistical strategies we use deal with the nonrandom sorting of teachers and students.

These are all important issues. How serious are they? I don't think they're terribly serious for policy research because I think a lot of these issues cancel themselves out with these huge data sets that we're working with. They're much more serious when we're talking about individual stakes for teachers.

Just quickly—and I'm going to wrap up in just one second—to give you an idea of how good these measures are, one indication of how good or bad they are, and it depends how you see this—these are three different studies that looked at the performance of teachers sort of pretenure and post-tenure. So the question is how good do they predict the longer-term performance of teachers?

And the number before the slash is teachers that were in the top

quintile before and were in the top quintile after. So we've got 46 percent, 29 percent, 22 to 32 percent, depending on the model. But you can see that somewhere between a third and close to a half of the teachers that were at the top pretenure were also at the top post-tenure.

Is this big or is this small? You know you are going to make some errors if you do this. Similarly, at the bottom, you see similar numbers at the bottom. If you're in the bottom quintile before, are you in the bottom quintile after? Clearly, it's not perfect prediction. Some people would say but it's not bad.

So what are the overall implications coming out of this? Number one, I think we should be using value-added measures freely for research purposes, trying to do the analysis as best we can, simultaneously trying to do the technical studies, trying to understand the weak underbellies of these measures. And if you go on the CALDER site, you can see both technical papers trying to get at these measures as well as substantive policy papers.

Individual, using for individual teachers. My own personal feeling is that this is important information and we don't want to suppress information. Do we want to use it mechanically? No, we don't want to use it mechanically. We know it's not perfect. We know there are problems with it. We want corroborating evidence, and there's lots of research just starting up now looking to see the extent to which value-added measures can be corroborated with things like principal judgment, judgment of peers, observation in the classroom, in order to get some corroboration, but my own personal feeling is this value-added information is good enough that it should

be on the table.

It should be on the table to help with professional development; it should be on the table to help with making decisions about who stays and who doesn't stay. And, yeah, there's probably going to be some error and who's going to bear the risks? The teacher or the kids?

If I could change one thing, and I'm speaking now as a researcher in the policy community right now, what I would want to do is move the tenure date out because if we have more data, we have more information and can make better decisions.

Right now a number of jurisdictions, teachers get tenure after 2 years. I think there are only a couple of states that go up to 4 years, but you could imagine if we could move this whole thing out to 5 years, really get over this learning curve part, really get some better estimates of the predictive validity of prior performance, then I think with, you know, initial selection and then what goes on in the schools after these teachers get hired, we could really boost the whole productivity of this system up.

MS. HOFFMAN: Thank you.

[Applause.]

MS. HOFFMAN: I'm not going to struggle with that one. We have, I think, two minutes for questions, and we've been asked if you do have a question, would you please use the microphone and give your name so that the transcriber can get the information down correctly?

MR. WHITE: I'm Andy White. I'm with the Center for Education Statistics, and I'm a research statistician.

For Dr. Hannaway, the teacher thing is old territory here, and I'm going to step right into the old part, but I kind of wish you wouldn't call them "worst." Maybe unproductive, inefficient.

DR. HANNAWAY: Okay.

MR. WHITE: Because worst means, you know, let's, where's the firing squad, I mean, and through these data sets, the real question is have you seen a way to find out what, in the first 3 or 4 years these underperforming teachers, is someone doing something? Like you showed one slide where they paired them with an effective teacher, and then it popped up. The person's performance popped up in the fourth year, I believe.

Can we see any, any on-the-job changes, any effort to find out if this person is just misplaced or is missing tools, should be a teacher or should be, or should get some help and then might be a teacher? Because we've got—this is a big job pool. Just because you shoot some of them doesn't mean that wading out in our job pool we've got effective teachers who don't know they're teachers yet.

So can the data help that?

DR. HANNAWAY: Well, let me take issue with something you started with, though, that this was sort of old hat. In fact, it's not so old hat. You know these are findings within the last basically 5 years.

MR. WHITE: My mistake.

DR. HANNAWAY: So, and then to your next point, what do we do about it, and I think that's where we're moving, what to do about it. First is identifying the issue, so the issue, you know, that's on the table is we have

this huge variation, and we don't know why. We can explain some of it with selection, but we're still going to have huge variation.

If you look at the data in New York, for example, if you look across, look at variation within New Teacher Project and within TFA and within other teachers, the variation within each of these groups is very large. So we clearly don't solve the whole problem, but in order to know, in order to be able to, number one, identify the teachers who are the least productive, we need time because they have to teach and do it, and then we need to get the data out and see.

I think Sean was talking about some of this, but you have to, you have the data in order to know who's doing it, and then we have to figure out what to do about it, and that's why I talked about if we could move that tenured-decision out, we could figure out who are the teachers who may have been slow at the opening gate and fast-learners, the one that, such as the one Sean talked about, or those that, you know, just can't hack it. This is tough work.

MR. MULVENON: If I could add one thing. Drafting off your point, Andy, a lot of people don't realize that like 85 percent of the school districts are rural so the idea of having an infinite population of teachers to replace those teachers is just not probable. So, you know, you're down on the delta. Say in Arkansas or Mississippi, you've got what you've got.

And the LDS Systems can be really fabulous in helping you in professional development programs to sustain that teaching pool and make them more effective which I think is a much bigger target obviously, but I

think this work is great to get us started on.

DR. HANNAWAY: Yeah, and we have to know how to do professional development. I mean right now—

MR. WHITE: That's the other IES grant program.

DR. HANNAWAY: Exactly. Yeah.

MR. KOEDINGER: Ken Koedinger, Carnegie Mellon University.

Jane, the Teach for America teachers, what do we know about how they're different from other teachers, you know, particularly about what they're doing in the classroom? But I also wonder about math SAT, for instance, in this data of math teaching effectiveness?

DR. HANNAWAY: Yeah. They tend to have stronger academic backgrounds. They tend to be inexperienced and they have stronger academic backgrounds. Teach for America in recent years has also sort of been building in various supports for its teachers, but I think—

MR. KOEDINGER: Does math SAT account for much of the effect?

DR. HANNAWAY: I'm sorry? We didn't try to disentangle what it was about TFA because we simply didn't have the information. I think TFA has a lot of information, you know, in its own data files that we just don't have, about how they go about selection and what are the personality characteristics and as well as the academic background characteristics.

The only thing that we can see from the data are, you know, scores on teacher tests. We know competitiveness of college. We know, you know, something about their academic background, whether they majored in

math and certified in math and that, but we can't tease the rest of it. We just don't have the data.

MS. HOFFMAN: One more gentleman has a question, and we're over time. So—

MR. SMITH: Robert Smith, Empirical Education. Thank you very much for a very nonideological discussion of things that are often quite heated.

Dr. Hannaway, a question. What about the practice of tenure itself? Do you see anything in your research about that practice?

DR. HANNAWAY: No. You know right now most teachers get tenure so there's little variation for us to study. If it were moved out and it were more of a decision point where more discrimination was used, we could probably do something from a research perspective, but at this point, we really can't say much about it except most people get it.

MS. HOFFMAN: Thank you. Thank you very much.

[Applause.]

[Whereupon, at 10:40 a.m., the panel session concluded.]