STATE OF FLORIDA
DEPARTMENT OF EDUCATION

AMERICAN INSTITUTES FOR RESEARCH

FLORIDA'S RACE TO THE TOP

STUDENT GROWTH IMPLEMENTATION

COMMITTEE MEETING

University of Central Florida

Teaching Academy Building

Orlando, Florida

Thursday, May 19, 2011

Volume 1

DEPARTMENT OF EDUCATION:
KATHY HEBDA, Deputy Chancellor for Educator Quality
JUAN COPA, Director, Research & Analysis

AIR MEMBERS PRESENT:
JON COHEN, Ph.D., Executive Vice-President
HAROLD DORAN, Ed.D., AIR, Principal Research Scientist
CHRISTY HOVANETZ
MARY ANN LEMKE

(Whereupon, the meeting was called to order by Kathy Hebda, after which the following occurred:)

* * * * * *

MS. HEBDA: Good morning, everyone, and welcome to the next meeting of the Student Growth Implementation Committee under Race to the Top. I'm Kathy Hebda, I'm the Deputy Chancellor for Educator Quality with the Department of Education. I'd like to not only welcome our committee members this morning but welcome our audience that might be watching over the web, and anybody present in the room. If there are audience members present in the room, we are very, very pleased that you're here. We'll have hard copies of the power point for you at the lunch break, and we also would remind you that this is a webcast and is always open to the public. We are very pleased that folks are interested in this meeting. This is a working group. The committee members, the ones that will be speaking during this time and doing the work of the committee, and we appreciate your participation.

Members, I'm going to let you begin by

introducing yourselves one more time. Most of you were here in person at the last meeting, but if you were participating from afar or not quite in the room, I'm going to let you do that. If you would just say your name and your position and who you represent, and we'll start with Ronda.

MS. BOURN: I'm Ronda Bourn. I'm supervisor of special projects at the Northeast Florida Educational Consortium.

MS. EDGECOMB: Doretha Edgecomb, school board member, Hillsborough County Schools.

MS. KRISHNAIYER: Latha Krishnaiyer, Broward County.

MR. MOREHOUSE: Lawrence Morehouse, president of Florida Education Department and professor of USF.

MR. LeTELLIER: John LeTellier, music teacher, Stanton Weirsdale Elementary School, Marion County.

MS. ACOSTA: Sandi Acosta. I'm a middle school science teacher at Kenwood KA Center in Miami.

MS. KEARSCHNER: I'm Linda Kearschner, I'm a business owner and I'm on the board of

directors for Florida PTA, and I'm here representing parents.

MS. WESTPHAL: Lori Westphal. I'm a teacher for hard of hearing at (inaudible) Early Learning Literacy Center, Lake County.

MR. CAMPUTARO: Joseph Camputaro, kindergarten teacher, Lee County schools.

MS. NOYA: Cristina Noya, St. Lucie County assistant principal.

MS. FRAKES: Stacey Frakes. I'm an instructional coach Madison County.

MR. FOERSTER: Sam Foerster, associate superintendent in Putnam County.

MS. FEILD: Gisela Feild, administrative director, Miami-Dade County.

MS. TOVINE: Gina Tovine, assistant superintendent Levy County.

MS. STEWART: Pam Stewart, deputy superintendent, St. Johns County.

MS. BROWN: Anna Brown, director for assessment of performance management, Hillsborough County Public Schools.

MS. WOODHOUSE-YOUNG: Tamar Woodhouse-Young. I'm a math teacher in Duval County.

MR. TOMEI: Lance Tomei. I'm the director for assessment accreditation in data management in the College of Education here at UCF.

MS. MARSALA: Nicole Marsala, 8th grade U.S. history teacher in Broward County.

MR. MURPHY: Jeff Murphy, director of student services Florida Virtual School.

MS. HALL: Stephanie Hall, Brevard County.

MS. HEBDA: Thank you very much, committee members. I appreciate that. Just for housekeeping purposes, was the sound okay? Could you hear everyone.

MR. ROBERTS: You need to be a little bit louder.

MS. HEBDA: Okay. Members, just for the folks that may be watching online, we do have the microphones placed around the table so if you could just direct your voice towards those microphones, it should pick it up fine.

I want to go through with you -- you have in your packet a power point that we will use just like we did last time when you were here that will kind of keep us framing your discussion and the process that you'll use throughout the next two days.

The first couple of things on that just to get everything set up before I turn it over to our experts, you have two sides that cover the meeting agenda for today and tomorrow. We begin at 8:30 each day and will adjourn at 5:00 today and at least by 5:00 tomorrow, no later than 5:00 tomorrow.

One of the things that I would say about the agenda, you can -- I won't read every line to you; you can see that on your own. Our contractors, the American Institutes for Research that worked with you the last time at the last meeting, what we've asked them to do this time for your benefit since your big goal this time is to work towards recommendations to the commission for a value added model to be used with FCAT data, is to provide you not just the results of all the data requests that you made of them at the last meeting but also a method for you to make and work towards that decision. And they developed a process that will help you do that and track the information that you receive from that data analysis and from the results of those data runs, and that will help you as work towards making your

decisions over the next two days. So I think that's something that you can look forward to and feel comfortable about because I know that you may be thinking, well, you asked for a lot of information last time and how are you going to receive that information and how will you make decisions about that information. And they have a process to help you do that.

So as I mentioned, the purpose of this committee -- just to remind everyone and everybody who might be watching -- the purpose of this committee is to make recommendations and the June 1st recommendations that you make, those that go to the Commissioner, you need to make a final selection by June 1st. That's really the first order of business that you have. It's not the last. June 1st is just a beginning. It's a process, the beginning of a process. It's a very important step, no doubt, but it is a first step and you'll be working throughout the remaining years of the grant to improve the models and recommendations that you make, look at how those models can be communicated, lots of things like that that you have on your agenda for the next three years.

So I know this work you're doing today is very important, and on behalf of the Commissioner, Chancellor, and the Department of Ed and State Board of Ed, I really want to say how much we appreciate your time and attention and your devotion to this task and the expertise that you bring with you, and just keeping in mind, though, that it is the beginning.

So here you are. You may remember these slides from last time. This is the process for sort of the year one of the grant with relationship to student growth. You already identified the initial models that you wanted and selected models for comparison. That's what you did at your last meeting and determining -- you had a discussion of the variables and business rules. One of the things we'll do after I finish these introductory slides is we're actually going to spend a couple of minutes reminding you of what those decisions were. In case you don't have your notes with you or any of those sorts of things, we'll lay out very succinctly where you've come so far. That will help remind you where you are and what you're going to do next, and of course, anybody

who's watching the proceedings, we'll also help them know exactly what's happened if they miss the first meeting.

So you finished all that at the last meeting. What's in the loop here is what you're going to do today and tomorrow. You're going to evaluate those selected models, the results of those things, how they run data using those models and the variables, and the business rules that you discussed; and you're going to compare those things. Remember, I told you we have a process that will help you the next two days to do that comparison in a way that you can feel comfortable about, there will be times for you to stop, reflect, always opportunities for questions and clarifications as you go throughout that process the next few days.

Then after the final selection is made by the Commissioner on June 1st then there will be reporting of results. School districts will get their data that was used to evaluate the model, the final model selected, data they can use to make local decisions about how they're going to use it in evaluation systems and their scales for ranking folks or rating individuals making decisions about performance next year. That's the process they'll be going through this summer after that data is provided to them.

So that's where you are in your timeline.

Again, this is the goal of the meeting. It's a simple goal; it's an important goal. Just reminding you again that your goal is to make recommendation. The Commissioner does have the responsibility to make the final selection. Then, of course, as that model is implemented next year and the following year and the year after that, every year there's a process built into the grant to make sure we analyze how effective the model was and ways to make improvements in the model as we go.

One other reminder before I turn it over to our AIR partners is their role throughout this process, just like you saw in the very first face-to-face meeting you had, their rule is not to make a recommendation. That's your role. Their role is to fulfill your request for data, provide information, answer your questions, lend expertise to the process, but the decision are, in fact, yours. I want to make sure that everybody is very, very clear about that.

Remember, I told you we're going to walk through the background of the decisions made so far and I want to bring everybody up to date, I'm going to call on Dr. Hovanetz to please do that part for us.

DR. HOVANETZ: Good morning. Just a brief reminder of where we've been and how we got to where we are today to look at the results that we have. A mere six weeks ago we were sitting here all together and we had our first face-to-face meeting where we actually narrowed it down to have AIR look at three selected models, a covariat model with fixed and random effects and a sustained differences model. We selected variables that we wanted them to include in the model and we selected business rules to guide the evaluation as well. We looked at and pulled out of Senate Bill 736 the three specific variables -- students with disabilities, English language learners, and attendance -- and defined each of those variables for inclusion in the evaluation.

We also had a webinar on April 14th where we did additional identification of variables to be included in the analysis. So we identified

gifted, age, class size, homogeneity of class, mobility, and school effects that will all appear in the models as well.

Just to reiterate some of the milestones, when initially presented we looked at eight different models. We selected three and at AIR we went ahead and evaluated those three models and some variants of those models. So you'll actually be seeing more results than for just three models today. I believe there are results for seven different models that we'll be looking at today. We provided guidance and direction. We'll show you how we incorporated the business rules that we made decisions on from last time. We'll talk specifically about how we identified and defined each one of the variables from the last meeting, and so today we'll be looking at all of these results making some initial recommendations to have it finalized by June 1st.

So I'd like to call Dr. Harold Doran up to start discussing through some of our initial results and findings.

DR. DORAN: Good morning, everybody. Thank you for having us back. So last time that I was

here I was really sick; this time I feel great. This time Jon is actually the sick one, so the coughing is coming from him not me.

This has been for us at AIR in working with DOE a really exciting six weeks. There have been a number of things that we really have an opportunity to look at, think about, analyze that we're going to bring to you today. So I think this is exciting.

I'd also like to say because I've been doing value added for a while and it's my opinion, which I think is a pretty substantiated opinion, that the work that this group is doing is the most comprehensive value added for comparison and in-depth analysis of different models that I know of that has ever happened to go into operation, at least people who have done the studies, and they thought about value added models. The depth of what this group has done has been very, very substantial and we're going to present a very substantial amount of results today.

Before I get started, I wanted to stop real quickly and look back at where we were and see if anybody had any questions that were pressing from the last time we were together, any particular concerns or issues, or if we're just ready to move on.

Hearing none, why don't we start? All right.

One of the commitments that we have in terms of implementing the value added models and coming back to this group is we want to be clear about what we call an estimand. What are the different value added models actually estimating in clear terms, because these are statistical models that are doing something that has to be understood in terms of their transparency, in terms of what they are estimating.

Essentially, there are two types of models that we're going to present today. We reviewed three genres of models last -- two genres of models, what we called the layered and persistence model or the learning path models and the covariate models. So now let's talk about what are the models actually estimating, the differences model is -- we say this, expect students who score the same in the pattern here to score the same and to continue to score the same, and assumes the same amount of growth for

each student in each achievement model. So students within performance level one are expected to have the same level of growth that all other students in that performance level. Students in the third performance level have a similar expectation.

I'll show you a graphic in just a moment. The covariate adjustment models, these models expect students who score the same in prior years to score the same the next year. Expected growth may vary within achievement level. It's that last part that makes these two models different.

In the first model, students within a similar performance category have similar growth expectations. In the latter, students in a similar performance category can have different growth expectations. That is the key difference between these two models.

Let's look at a visual display. Sometimes it's good to look at the human language interpretation of what the models are actually estimating and some others like to look at some of the plots. I'm going to try to present information in multiple ways today so that it

really saturates and we understand what we're estimating.

The green line is the line that would be fit through a covariate adjustment model, and essentially, one of the things that we see here in this red line is the simple differences model. It essentially has a slope of one within each of the performance categories, whereas the other model has a slope of whatever it is. It's not necessarily one; it doesn't have to be constrained to be one. But it's not the same within a performance category where you see the red line expects lower growth for students at the lower end of the score distribution relative to the covariate adjustment model. The important point that we want this group to recognize is that the models differ in terms of their expectations for growth for students within a particular performance model. My team want to add anything?

DR. COHEN: Yes, I want to add a little bit. This is -- we've boiled it down to just two classes of models earlier.

DR. DORAN: The web books have -- the web books have these, is that correct.

MS. HEBDA: Is that correct.

DR. COHEN: Just to make sure that everyone followed, the X axis is the kids' score last year; the Y axis is the kids' score this year. The little blue dots -- okay. We've plotted all the kids in grade seven, these are their math scores. We chose a grade as an example so you can see the true relationship. The covariate adjustment model gives you the best fitting line if you were to fit that. So it minimizes the amount of dispersion around that prediction. Does that make sense?

Every kid who got -- and the echo models have some other covariates in it -- but as I said a little loosely, but the general idea is that every kid who got a 1500, you go up to the covariate adjustment model, they would land on that green line, come over and you might expect 1600; it comes out to about there. So every kid has the same expectation if you have the same scoreboard.

Does everyone understand what the green line is?

The red line, it's just a simple difference. All we did was we said we're going to look at your score and subtract out the average growth from within that group. We'll just look at the average growth for kids who started at the same achievement level; and instead of having a best fitting slope, it has a slope of one. You look, it's hard for you to see in some places where these lines differ. But you do see that they cross a number of times. Where the red line is above the green line, that means that you're predicting higher performance among the kids than is typically observed. The green line is what's typically observed; it's also the line hit by the covariate adjustment model.

So down here the difference in models predicting less growth than is typically observed. That would tend to say that teachers teaching these students would typically exceed that more readily. At the other end you see this is where it goes if you follow this, at the end of this group, teachers teaching these kids would have a harder time exceeding that because you're expecting more growth than is typically observed. Does that make sense? That's the difference between the covariate adjustment

model and simple differences model. So we're clear? Questions? Okay, back to Harold.

DR. DORAN: All right. Let's go to the next slide. So let's talk about a summary of the models that have been estimated. Let me talk through this and then I'm going to give you something that you're going to keep next to you throughout today and tomorrow. The first model, what we call Model 1, we call it a two-level model -- a two-level model that includes teacher effects, controlled for one year prior achievements, we call those lags -- so one year prior achievement is called one lag -- and control variables that are ELL or English language learners, SWD or students with disabilities, and attendance, and we estimate teacher effects as random effects. I'm going to talk about random effects in just a moment. I've got a reminder slide in here in terms of what they are.

We call it a two-level model because the data are students at the first level, teachers at the second level. That's the terminology that we use here. Okay. That's Model 1.

Model 1A is the same as Model 1 except that

we have two years of prior achievement data. What I mean by that is in this regression we have what's called a dependent variable -- that's the current year's score -- and then we use two prior years of achievement data. We call those lags. So one lag means in the regression model, one of our independent variables is a prior test score. If we had two lags then we're using two prior test scores. The rationale for that and whether or not that matters is going to become clear as we look at some of the results comparing Model 1 and Model 1A.

Model 2. It's the same as Model 1 but estimated with fixed effects. Let me say something now and we'll talk more about this throughout the day if we need to. We estimated the model with fixed effects just as we said we would. We're not presenting them today although we can fully talk about them. Let me explain why.

Algebraically, mathematically, we know that the random effects and the fixed effects are guaranteed to be the same as the number of kids in a class gets larger. That's the constraint

that has to be in place. They can be the same and they estimate the same quantity. Early on in the analysis we saw that the models with fixed effects were yielding unstable results. We spent quite a bit of time looking into this and we know why. There is a business rule -- there were a couple of issues. There is a rule that allows for students, I forget exactly how to say this -- Christy or Jon, remind me -- students who are taught by multiple teachers to be partitioned across these multiple teachers. Essentially what was happening is there's teacher one who has a group of kids, teacher two who has the exact same group of kids, okay? Two different teachers, exact same group of kids.

In a fixed effect model you can't have that, all right? The term we use is called co-linearity. We have to remove one of those teachers. We're estimating the same exact thing. In fact, you can remove one of those teachers because they're duplicated without any consequence on model estimation.

There's another condition happening in the data which causes near co-linearity. They're not exactly the same kids but they're pretty close to it. It would almost have the exact same number of kids but they might be off by one kid or two. That causes for a problem in the mathematical estimation of the model. What was happening was that those teacher effects were causing -- those teachers in the data, without conditions it appears to be true -- were causing for the results to be unstable. In fact, if we remove some of those issues and we estimate the model with fixed effects and we correlate it with the random effects, they are correlated better than 0.9. A correlation ranges from –1 to 1; a correlation of 1 means there is a perfect relationship between the two, a 1 to 1 correspondence. A correlation of zero means there is no correspondence between the two estimates. The closer you get to 1, the greater that correspondence between those two is. A correlation better than 0.9 tells us what we hypothesize about this. When we were here last time six weeks ago that the models were estimating the same thing and that turns out to be true in the data.

The issue here is and the reason we're not presenting it is because there is a business

rule that prevents these models from being estimated and presented in a reasonable way to this group, but we do know that we could rely on the fact that both algebraically and in the real world the results are highly correlated when we remove that particular issue and when we don't take it into consideration.

Do you want to add anything?

DR. COHEN: No, that's basically it. You wind up having to toss out a lot of teachers because of fixed effects model; it's not true with the random effects model. You don't want to really toss out a lot of the teachers, you don't really want unstable estimates.

DR. DORAN: All right. Model 3 and variance. A three-level model that includes teacher and school effects. Control for two prior achievement -- two years of prior achievement, two lags, and varies as to which variables are included. That's the general class of Model 3. So we have teacher and school effects. It's a three-level model because the data are structured in this -- we have students, teachers, and schools. That's the terminology that we use. We have two-level models and the

three-level model.

Model 3A has no additional variance. That's it. The description that we see.

Then we have Model 3A1. Model 3A1 is Model 3A but it differs only in terms of the number of prior achievement test scores. It has one lag, not two.

Model 3B. This is the description, but it includes ELL, SWD, and attendance. We use two prior test scores -- oh, there's a note there. It always uses two prior test scores.

Model 3c, ELL, SWD, attendance, and the following additional variable -- class size, homogeneity of class composition. Let me explain that variable. That is a variable that describes how similar students are within a class, all right. So we construct a variable, call it the homogeneity variable, and it's essentially -- I think we've got another slide that describes it, but I'm going to mention this now. I have to find the easel to talk about things multiple times. It is the difference between the students at the 75th percentile within a class and the 25th percentile within a class. So if students within a class are very

similar in terms of their prior test scores, that difference will be small.

If students within a class, if a teacher has a class that is very different in terms of their ability, that number will be very big. It's a spread. It's essentially a spread between kids within class. How different are kids within a class.

Mobility, student mobility, and difference from modal H. Christy will describe those in just a moment when we get to those slides. And we use one or two years of prior achievement data -- here again, this one uses two prior achievement scores, okay.

And last, Model 4 is the differences model. All right. This is the language description, the narrative description of what these models are doing. In your folders you have this -- can everybody pull this out to make sure we all have this here? I'm going to refer to this as the scorecard throughout the day, today and tomorrow.

This is the same as what you see here, but you're going to want to keep this next to you throughout the day. We're going to talk about the models and all of our slides use labels like 1, 1A, 3A, 3A1, and so forth, and we don't want you to get lost in the details of the models. So what we've created is this little scorecard; and essentially what you see is Model 1 and its key characteristics. So, for example, Model 1 has one lag and had teacher effects only. No school effects. The covariates that are included are SWD, ELL, attendance, and the effects, the teacher effects are random, and school effects if they are included. There's no school effects in Model 1.

All right. Now I'm calling this a scorecard because one of the things that we're hoping we will do is by the end of the day we have to make a recommendation, and Sam is going to facilitate a conversation. What we have over here in terms of notes, these are the primary evaluation criteria by which we'll be looking at these models today. We have data that show you how these models stack up against each other based on those criteria.

One of the things we're hoping this scorecard will be used for is as we talk about precision, for example, and the precision of the

different models, you will come up with your own ranking system and maybe you'll use a 1 to 10 scale. Maybe you'll use a 1 to 50 scale. Maybe you like happy faces, pluses, minuses, whatever is comfortable for you, and if you like a model, for example, and you're using -- you might put for that particular model, Model 1, a precision plus or an 'A'; or if you don't like it, you might put an 'F' based on the data and the results. And essentially you might use that to have a ranking for each model on your own personal scale for precision. Then include school effects. Do you like to include school effects? And so forth.

By the time we get to Sam and he facilitates the conversation about whether or not you're ready to make a decision or intend to make a decision on a model, you'll be able to look at this and you'll look across the rows, and you'll say, well, I look at Model 1 and I have all the sad faces. They're all minuses there. In my own opinion in these criteria, I can eliminate that model from my choice. I look at Model 3B. I have all those pluses there. Then your decision by the time you get to the

conversation with Sam is not random.

Have you forgotten anything? You've evaluated the models based on empirical criteria. There's going to be a ton of discussion and other things that you care about perhaps beyond this, but the goal here is we have to have a process by which we evaluate the models that's better than our opinions on what we think the world should look like. So we made an attempt not only to estimate a large number of models with multiple areas, but to also come up with what we believe are reasonable indicators that you can use as the lens by which you can evaluate the model, then this decision is yours. It's not us standing here telling you this model needs to compute, it's statistically very nice. It looks good, the plot looks good, and we remove hopefully all of it. Keep this by your side throughout the day.

Jon, go ahead.

DR. COHEN: I just want to help with a little bit of organization here. It's not entirely clear from the chart -- I mean, it's there but it doesn't just pop out at you -- Model 1 doesn't include school effects. It's

only teacher effects. Model 3 includes teacher and school effects. We've tried to have at least pairs of models so you can evaluate different decisions, and essentially there are four -- these models vary along four dimensions. One dimension we've already talked about. Covariate model versus differences model. That would be Model 4 versus all the other models. Model 4 is the differences model that we just talked about; all the other models are covariate models.

The second dimension is whether you include school effects in addition to teacher effects; whether you're simultaneously estimating school effects and Model 1 does not include a school effects estimate, Model 3 does.

The next decision is how many years of prior achievement do you include? We recorded those as lags and you'll see them in the first two columns, so we have one teacher effect only and one -- we have two lags and one lag in each Model 3 and Model 1, if you want to look at those.

Finally, the question is which covariates do you include? And we have everything that

ranges from none to a few to the kitchen sink.
MR. TOMEI: Quick question. I'm looking at the chart and it appears that the lags for 3A and 3A1 may be inverted, based on what's on the slide; is that correct?
DR. COHEN: Right, 3A and 3A1 are inverted, yes, thank you for that. So 3A should have a checkmark under two lags; 3A1 should have a checkmark under one lag. Thank you for that. That's a good catch.
DR. DORAN: Let's all make that change to make sure we're all on the same page.
MS. FEILD: Shouldn't 3B have the covariate ELL, SWD, and attendance based on that?
DR. DORAN: Yes. Thank you.
DR. COHEN: Excellent. Thank you.
COMMITTEE MEMBER: Would you say that again, please?
DR. DORAN: Yeah, SWD under row 3B under covariates, in 3B under covariates, write SWD, ELL, and attendance. Under Model 3A in that row, remove the checkmark under one lag and instead put the checkmark under two lags. In the row below it, Model 3A1, put a checkmark in the column for one lag and remove the column for

two lags. Thank you for that catch; I'm sorry for the error.
DR. COHEN: Okay, so there are four dimensions to the models. Differences versus covariate, four versus everybody else. Only teacher effects and school effects, Model 1 versus Model 3, the number of prior year achievement -- those are called lags -- and the particular covariates. So these are the four dimensions you want to be thinking about.
DR. DORAN: Yes, go ahead?
MS. FEILD: I have a general question. When we talk about your data, are we only talking about students that have been promoted? In other words, if you're looking at three years of data, would that be progression from 3rd grade to 4th grade to 5th grade, or would we be including data 4th grader retained, 4th grade, 5th grade?
DR. DORAN: Christy, how did we do retained students? Did we have -- if we had, say, a 5th grade student who was retained, would we be using the retained data -- would we use two prior years of data or would we only use the data if they were promoted in this sequential

order?
DR. HOVANETZ: We would use the prior years.
DR. DORAN: Any scores.
MS. FEILD: Regardless of the grade level.
DR. DORAN: I don't remember that -- (inaudible) -- if we use any prior school --
DR. HOVANETZ: With the exception of 3rd graders, we do not use any 3rd graders.
MS. FEILD: What about retained 3rd graders? No?
DR. DORAN: No. Remember, we cannot estimate teacher effects in 3rd grade because there's no prior achievement data. So the only -- you have students who have two 3rd grade scores because they were retained for some reason, you'd get biased effects because of that.
MS. BROWN: On the chart 3C in the covariate, it lists gifted but I don't see gifted on the slide as a covariate.
DR. DORAN: Gifted is in the model. It should be on the slide.
DR. COHEN: It's actually in every model that includes SWD, it also includes gifted.

DR. DORAN: Let's add that. Any place you see SWD, also add gifted.
DR. COHEN: It should say exceptional students.
DR. DORAN: I was so certain I had this right.
MS. EDGECOMB: I want to make sure that I understand our responsibility under the section called Notes where you talked about using some kind of quoting system, some smiley faces, the plus and minuses. Do we each do these individually? Is that correct?
DR. DORAN: That's right.
MS. EDGECOMB: What will help the process so that we limit subjectivity in doing this?
DR. DORAN: Good question. When we talk about precision, we're going to show what we need by precision and we're going to show you results. We'll actually show you how the models vary in the terms of their precision. We'll find precision. Then we'll show which models are more precise and less precise, and then you'll be able to make your judgment based on data. Which of the models are more or less precise? Each of these criteria are associated

with data, so there's no virtually no subjectivity in that regard. Whether or not the models are precise is an empirical question. Whether or not they are school effects, we show you the consequences of school effects or not, and there's data. Same thing with parsimony, classification, accuracy, and lags. There are data that we will present associated with each of these criteria and you can make a judgment on whether Model 1 is better than Model 2 based on the results of what we show you.

Christy?

DR. HOVANETZ: Just for clarification, this is just an advance organizer for you all to take into -- we're not necessarily going to ask you to keep track of points or numbers, but just for you to be able to reflect in an organized way and that we're all doing it the same way. So when we're talking about Model 3A, you can look at 3A and see the notes that you've taken for that specific model to help refresh your memory because there are seven different models that have not -- very fancy names.

DR. DORAN: Yes, ma'am?

MS. BROWN: I just want to clarify the

gifted SWD. Was gifted included as a separate covariate in each of these situations, in addition to SWD, or included as SWD?

DR. DORAN: Included as a separate covariate.

MS. BROWN: Thank you.

DR. DORAN: Yes?

MR. LE TELLIER: Under, for example, precision, what you said is totally objective.

DR. DORAN: Virtually.

MR. LE TELLIER: Or virtually. If you come out there with the data because -- especially with its title precision, is there an absolute precise way for all of us to know this is the model that was the most precise versus us guessing it?

DR. DORAN: Yes. We're going to show you -- we're going to define precision, we're going to talk about what are we looking for when we talk about precision. That's a question. Then we're going to say what's the statistic that tells us what defines precision, what statistic defines precision. Then we've got criteria. How do you know a model is more precise than another? What are we looking for in that

statistic? Then there's a why. Why do you care about this? So each of these criteria will be structured around those four things: What's the question? What's the statistic that answers that question? What's the evidence -- what are we looking for in that statistic? And then why do you care about this?

Yes.

MR. TOMEI: Kind of related to the question on gifted, within SWD how many different categories did you look at independently?

DR. DORAN: We looked at -- Christy, remind me.

DR. HOVANETZ: We going to talk about this in two slides.

DR. DORAN: Yeah, right, we're just about to transition over --

MR. TOMEI: We know from research in other areas, we know there are differences among the different --

DR. DORAN: Yes.

DR. HOVANETZ: There are --

DR. DORAN: I'll get to this and then I'll pass it over to Christy for the actual operation lies -- where the variables are.

I'm going to actually get a glance -- this is a slide we talked about last time. I already mentioned this when I was talking about the differences between the fixed and random effects earlier. But I want you to recall that fixed and random effects are different ways of estimating the same thing. In fact, mathematically we know that they're expected to be the same. They're expected to be no different from each other under the condition that classes get bigger. You have a lot of kids in a class, then the numbers start to look about the same. How many? I don't know exactly, 20 or 25 or 30 kids in a class; those numbers should convert to the same value.

The reason we estimate these different models is because there are different classes of statisticians, some people who like to treat these models as fixed effects, some who like to estimate the most random effects. There are statistical nuances, there are certain properties of the random effects that people like in certain properties of the fixed effects that people like, and essentially what we've shown is that when we remove those teachers who

are near colineal, the models are essentially doing what we hypothesized; they're estimating the same thing. There is essentially no difference in what is being estimated.
And when I talk about what, I'm talking about this thing called a teacher effect, value-added effect.
Christy, why don't I toss this over to you to do what's next?
DR. HOVANETZ: We'll update the scorecard or note taking device and make copies of it so you'll have a new one.
Since we already had started talking about the variable discussion, just to be very clear why we are talking about including variables in the models to begin with; and just to refresh our memory as to the discussions that we had on April 4th and 5th and again on the 14th at our webinar, the reason we are looking at adding controlled variables is to reduce the variances' unequal distribution of students that (inaudible) in teachers' courses. There's limited debate. We had this conversation back on the 4th and 5th and we did on the 14th about whether or not adding a lot of controlled

variables is going to make a difference in the model, whether or not it's going to make it more precise and whether or not it's going to actually level the playing field for teachers.
Some of the rationales for including student characteristics is to eliminate that bias, but the policy implications of it is student who want to set it for differing expectations for different students. So just a little reminder or refresher about the conversation that we were having on these variables before.
Reminder on the framework that we operated under when we were talking about which variables to include. First of all, we went through and we looked at the variables that were in Senate Bill 736, the SWD, the ELL, and the attendance, and then we had the brainstorming session where you all listed out 20 or so variables that we had conversations about initially on the 4th and 5th and then went through each of them in detail on the 14th and made judgments about them then. When we talked about including variables in the evaluation, we talked about whether or not it was a variable that was within the teacher's

control, whether or not it was already being measured by another variable that we were looking at, and also whether or not it could be explained by pre-test data. That was the framework that we were operating under to put variables into the model for evaluation.
We evaluated a lot of variables where we thought let's just see what this looks like and we'll base our judgments on the results, but we will be using the same framework as we're considering the results of these variables, not just whether or not they're significant but whether or not it does make a difference in the precision of the model and whether or not policy-wise it's the right variable to be included. So just keep those same conversations in mind that we had before.
Okay. This is the list of variables that have been evaluated within the models. Students with disabilities status was done with a dichotomous variable for each of the individual disabilities. So we can either play a game where you guess which variables are D, E, Z, or I can just tell you. We did not include the exceptionality codes listed here based on the

recommendations of this group last time. So as a reminder, occupational and physical therapy were not included as variables. Z is a not applicable variable; U is a established conditions; T is developmentally delayed; M is hospital or homebound; C is orthopedically impaired; F is speech impaired; and L is gifted. Those are not considered disability categories that we evaluated within these models.
We had 14 different SWD exceptionalities that we did evaluate within the model. Four of these exceptionality codes have been collapsed and so we will only be presenting on ten of these exceptionalities. What happened is there were three codes that were collapsed into the W code, which is intellectual disability, back in 2007-08; so you won't see as many disabilities or exceptionality codes as we talked about because they don't exist in the data anymore. So that's where you'll see a difference in what we talked about versus what's being presented.
Gifted status was done as it's own independence variable. Dichotomous variable. The student was listed as gifted or not listed as gifted. With the students with disabilities

codes, we had talked about looking at primary and other exceptionalities. This is only looking at primary student disability. We did not look at secondary or other exceptionalities. It's only primary.

For English language learner status, this is also a dichotomous variable. Students have to be coded as LY or currently receiving ESL services, and they can only be coded as LY for two or fewer years. So if a student is in his third year of receiving services, they are not considered an ELL student for our purposes. Yes.

MS. ACOSTA: Just to clarify, the LY classification has really no bearing on their ESOL level, so in other words they could have been receiving for two years and have reached ESOL level four or still be in ESOL one, and they'll be treated equally?

DR. HOVANETZ: Yes.

For attendance, we treated attendance as a continuous variable the number of days the student was in attendance at the school. So if a student was in multiple schools, we added the number of days in attendance in all of those

schools to get the total number of days the student was in attendance. It's a continuous variable, just the number of days present. Just as a reminder for this particular variable, this is information that comes in during survey five which is end of the summer and so for the evaluations we've been able to include this in the simulations that we've been doing, but this is going to be a data capture that will work with the Department if we choose to include it to be sure that we can get timely information on attendance.

Class size is a continuous variable. We did not use state determined class sizes. We actually just counted the number of students enrolled in a course and that was the number of students or the class size for that particular course.

Homogeneity of class composition. Harold mentioned this already, but in order to determine the homogeneity of a class we looked at the score at the 25th percentile and the 75th percentile and took the difference. If it's a small difference between the 25th and 75th percentile, it's a homogenous or more homogenous

class, a larger difference is a less homogenous class and this was done on a continuous variable.

The mobility calculation, also a continuous variable, we looked at the number of transitions a student made from school to school. If we have only one record for the student, the student had zero transitions. If we have two records for the students in different schools, that's considered one transition. So each time a student changed schools during the school year, it's counted as a transition. We did encounter some students that had two records with two entry dates into the same school. If there was a 21-day period between the exit date of the school and the following entry date into that same school, they were considered to have made a transition. If they spent time somewhere else, it may not have been in Florida, but we're not at that school that actually made transitions.

Age is also a continuous variable. We looked at and calculated the modal age for the grade as of September 1st and took the difference between the modal age and the

student's actual age to come up with the age difference.

And I will turn it back to Dr. Doran.

DR. DORAN: All right.

DR. HOVANETZ: Well, first, do you have any questions about variables?

MR. MOREHOUSE: I have a question about how you define homogeneity. Is it devised strictly on test scores?

DR. HOVANETZ: Yes, strictly on prior student achievement.

MR. MOREHOUSE: And what do you mean by "unique course" or "each unique course"?

DR. HOVANETZ: For a unique course, we looked at the district and the school, the courses offered, the course number, the period the course was offered, and the teacher for which the course was offered. So a unique course is any course number that is unique by school and by district and period.

Lance.

MR. TOMEI: I have a question about the class size statistic. How are the data collected and how stable is that particular statistic between data collection points?

DR. HOVANETZ: We are only looking at survey two data and survey three data, so it's that second week in October and that second week in February, and it's whichever students are enrolled during those two specific weeks in that particular course in that school, in that district with that period number. That's considered the number or the count of enrollment for class size.

MR. TOMEI: I'd like to ask the P-12 reps on the committee. Do you see that as a fairly stable statistic or is that one where there could be variance in the statistic itself that's not going to be captured in the reported data that will be used in the model?

MS. FEILD: Why not -- it was sort of a combination question. First, what was the self count within the class period teacher that was used to either aggregate the data or not? Secondly, how are we handling semester courses in the high schools? Generally, the kids are FTE's, did we account for that?

DR. HOVANETZ: Yes and no. So unfortunately we don't have a minimum class size to a discussion that we can have, but based on

the estimations that we did for these particular models, we did not limit the number of students that had to be in a course in order for a teacher to generate a value-added score.

With respect to semester versus full year courses, we did not look at whether or not it was a half-year course or a full year course. The data that we received showed us if a student was enrolled during survey two or survey three, if they were enrolled in the same course in survey three that they were enrolled in survey two, that course was eliminated. So the data information that we have shows each of the students' enrollments in the school in that particular course.

MS. FEILD: So the data polled for the courses, what was exactly polled? Are we polling only reading and math courses traditionally?

DR. HOVANETZ: Yes, we are polling only courses that are listed in the course code directory as a reading course and/or as a math code.

MS. FEILD: And does that include the ESL, reading through ESL and ESE?

DR. HOVANETZ: I have an entire course code directory that I can show you there are 166 courses that were polled for reading and 90 courses that were polled for math, and it's based on the Department's determination in the course code directory of what is a math course and what is listed as a reading or English language arts course. And in your packet it shows specifically there's a course code directory on how those determinations were made by the Department. It's a separate handout that was behind your Power Point on the right-hand side and the reading courses are courses that are identified as requiring a reading certification to teach them or a course that is mandated by the State Board of Education as a remedial course, a remedial reading course.

Math courses are identified by the prefix, those are a little bit more simplistic to identify. English language arts was determined by a committee and then also by the course code prefix. So I have a list of all the courses that were included for the particular analysis.

MS. YOUNG: I remember the discussion about attendance, but I didn't remember the outcome.

For students that are in multiple courses during the day and they have a tendency to leave after lunch, did we -- do we have data for that or just the whole day or they're marked present for the whole day, that's it?

DR. HOVANETZ: Yes. There was no way to capture by course at this time.

MR. COPA: Christy mentioned this but just one thing to add. As we talked last time in April, we have limitations on what we currently have to model, but one of the outcomes of one out of this process is to look at different ways to capture the data, improvements in the data systems, and we have already started conversations at the department level about capturing a lot of this information such as attendance at a course level as opposed to the daily attendance that's currently collected.

MS. FEILD: Referring back to Lance's question, I think there will be issues. Probably not monumental, but I know there are districts that offer some of the core courses within a semester. For example, algebra which is not going to be dealing with an end of course; a lot of districts are planning on offering it as a semester so students can request -- . So we'll have to take that into account.

DR. HOVANETZ: And that's okay, though, because we will still have the district, the school, and the course number for that particular student. So we do have that information right now and in the evaluation.

Other questions? Ask as many questions as you want today. Today is all about just getting you information, making sure that you're comfortable with the results based on the models we've selected or you selected, and we want to get all the information because we'll start making decisions tomorrow, or you'll all be starting to make decisions tomorrow about which ones we're leaning towards.

MS. EDGECOMB: In one of the earlier slides when you began talking this morning, you talked about the importance of data availability and accuracy. Is the assumption that all districts have in place the capacity to provide those two characteristics about data and input?

DR. HOVANETZ: That's a great question. I think a lot of the data that we are using is

accurate because it's been used for other purposes. There are some pieces of information that will hopefully get more accurate as we continue to use them, but this is the best available information and as we select which variables to be included and highlight that with the districts that there will be more attention paid to insuring the accuracy of those. We'll also have quality check processes in place where districts are signing off on their data, just as they do with other school accountability measures for school grades and AYP, so they'll have the opportunity to review the data before these results come out. But it's a great question.

MS. EDGECOMB: Okay. Can I continue just a little bit with that? Is there going to be a timeline when all the districts do or perhaps for those districts who aren't up to par that will give them the opportunity to develop some systems that, data collection or dashboards or whatever, to get those things in place knowing that these are the expectations?

DR. HOVANETZ: Sure. As Juan has mentioned, they are in the process of revising

how data is currently being collected and putting in new data collection procedures. All the information we're reporting here is information that has been collected for decades. So they have a system in place to capture this data information that we talked specifically about here. I think those need to be refined. The department will be working on that to assist districts and process these in order to be sure that is all reflected accurately.

MS. FEILD: So let me make sure I understand the premise of the analysis, so an elementary teacher who is sitting in a classroom with 25 kids; at the end of the year you'll look at the data for the 25 kids. You'll see whether those kids were there for both FTE's. If only 20 of them were there for both FTE's, those 20 will comprise her data analysis, correct? Is that correct?

DR. HOVANETZ: If the student was enrolled in survey two or three, they're included in the analysis right now.

MS. FEILD: "Or" did you say or "and"?

DR. HOVANETZ: Right now, it's "or". So the information that we have is for --

MS. FEILD: I thought you said "and" at the beginning.

DR. HOVANETZ: Or. We have survey two information and survey three information. The data set that we have has students that were enrolled in survey two and enrolled in survey three. If their survey three records is identical to their survey two record, the survey two record was removed from our data study. So if they're in survey two and not in survey three, the assumption is that they were there for a majority of the year. Or that's the assumption we made based on the data. So if they are enrolled in any of the courses with that teacher, they're accountable with that particular teacher.

MS. FEILD: So a child who switches schools will be counted in both schools?

DR. HOVANETZ: Correct.

MS. FEILD: So he's in teacher A at the beginning of the year, they move and in January they move to a different school with a different teacher?

DR. HOVANETZ: Correct.

MS. FEILD: And the driver here was the

course not the teacher, correct? We were talking here about reading endorsement, but if a teacher is a social studies teacher in a high school who happens to have a reading endorsement and is teaching two periods of intensive reading and she has 25 kids, which we have now a lot -- right, 25 kids in each course, tenth grade for periods of European history, two periods of intensive reading, she will have data for those 50 kids and those two intensive reading courses?

DR. HOVANETZ: Yes, this is based on the courses the student are holding.

MS. FEILD: I could go on. I could ask another question. Elementary school, self-contained teachers who are teaching both reading and math would have two value-added scores, a reading score and a math score?

DR. HOVANETZ: Correct.

MS. FEILD: Assuming the school coded that as a self-contained?

DR. HOVANETZ: Correct.

MS. FEILD: So if in fact the school is departmentalizing it but they did not code the teacher as departmentalized, which we know they do, then the teacher will be attributed math

scores in essence and in fact that teacher didn't really have math scores.

DR. HOVANETZ: Correct.

MS. FEILD: I think this alludes to Miss Doretha's comment on accurate scheduling and data pieces in the past; that was not important.

DR. HOVANETZ: Correct. We don't have actual runs of which teacher was assigned reading and which teacher was assigned math by a particular teacher, but one of the things we found is an elementary school teacher was more likely to have multiple teachers than a middle school or high school teacher was. Elementary schools are doing a good job of parsing out the students taking spelling, the students taking writing, the students taking reading, the students taking math; and so they have multiple course enrollments, it's not just fourth grade. So that's been a big shift in the data over the last five or ten years. But we are seeing that more students in elementary school have multiple courses associated with them.

MR. COPA: Just another disclaimer again. These are all great questions, and again back to the issue that we're operating under the

constraints of what we have currently collected for this modeling purpose, but just so everyone is aware as well, both the law requires a roster verification process to be in place for this when this is operational, the Department with partner districts. Hillsborough is one, NEFEC is another one, and also Osceola County of developing a teacher-student data link roster verification system through a grant process with SELT, which is -- I don't know what -- I can't think of what the acronym is right now, but they have a grant through the Gates Foundation. So we're working with them over the next year and coming months and we'll be putting forth a process in place working with our district partners and open to the entire state on a roster verification system to improve that data that will be so fundamental to this purpose so that we can deal with those issues such as the variability in how schools or districts may report this course information, since it hasn't been used for this high stakes accountability purpose in the past.

MR. LE TELLIER: Going on with what you were saying with the count two and three if the

student moves from one school to another, is the teacher that had them for the second count or the October -- would that affect their value-added model at the number three?

DR. HOVANETZ: That's a great question. That goes to the growth expectations and attributions, which is in a few slides, but maybe I'll skip to it now just to address your specific rule.

This is important for just kind of the fundamental understanding of how we determined expectations. Students' growth expectations are determined on the courses they are enrolled in, so if the student is enrolled in a reading course their expectation is based on that reading course. If a student is enrolled in multiple reading courses, their expectation is based on multiple course enrollments regardless of which teachers they had. It's based on the number of courses the student has taken in that subject to develop the student growth expectation. Harold will talk a little bit more about what those expectations are, but essentially we looked at students enrolled in 1 to 6 reading courses, 1 to 6 math courses and found the expectations of students in two or more reading courses higher than a student enrolled in just a reading course, but the difference between a student enrolled in 2 and 3 reading courses was not significant. So the expectation of growth for a student is based on the number of courses they're enrolled in.

The attribution of that growth is given to each teacher that student had. So if the student's growth expectation is based on one course and they've had one teacher, that teacher is fully accountable for that student's growth. If a student is enrolled in two or more courses, that growth expectation is a little bit higher and that teacher -- both of those teachers that that student had -- whether it's the same teacher or different teachers, both of those teachers are accountable for that higher growth expectation.

So if Gisela and I both had reading courses -- she was teaching one reading course and I was teaching another and Stephanie was in our class, she has a higher growth expectation because she is taking two courses. I'm fully accountable for what you do in meeting that higher growth

expectation and Gisela is fully accountable for you meeting that higher growth expectation. So we're both accountable for that higher expectation.

MS. FEILD: So how would the homogeneity of the class composition -- how does that work into a double portion? You'll have a lot of kids who'll have English and intensive reading. So how does that variable work?

DR. HOVANETZ: The homogeneity is a course level variable and it's generated for each particular course and it's used as a controlled variable as a model.

DR. COHEN: So we drove right into the fine details. So there's a difference between a course and a class. I'm a student in school A and move halfway through the year to school B, I'm taking Algebra 1 in both schools; that's one course. It's two classes. The homogeneity variable is calculated for each class separately. So if I was in a very broad diverse -- diverse in terms of kids entering math skills, in the first school I would have one homogeneity variable that had a large number. That is not very homogenous. Then I move to a

school and I'm in a class and it's all kids who are struggling with algebra just like I am, it would have very little homogeneity. I have both those variables in the model. So that's a class level as opposed to a course level model.

I want to remind you all, you guys are the working group and the committee, so the attribution that Christy is talking about is what was understood to come out of the committee before. There's nothing that stops you from revisiting that. We're here to implement your affectations. So mathematically -- let's think about the easiest case. You're a teacher who's teaching just a team taught course and there are two teachers in the classroom with the same kids, and that's the only class you teach. Let's say you're an elementary teacher; it's the only class you teach. It doesn't matter whether it's a hundred percent or fifty percent attributable to you because all of it is like a weighted average, right? So the average multiplying everything by fifty percent, each kid by fifty percent, it's the same as the average is going to be if you multiplied each kid by one. Does that make sense?

The only time it makes a difference is when you're teaching kids with a differential number of courses. So I have some of my kids who I'm their only teacher for, they count one; I have some kids who are in my class and in another reading class. Then do they count one or when I calculate my average do they count as one kid or do they count half as much as the other kids because I would only be half-teaching them? The decision that we understood from the last meeting was that all kids should count equally no matter how many other teachers are teaching them. That was the attribution that Christy was talking about.

Are there questions about that?

MR. LeTELLIER: I don't think I quite got what I was trying to say, my point, across. Survey 2, survey 3. Is that teacher that had them at the survey 2 point accountable for what the survey 3 teacher where they get them to? Let's say that at the survey 2 point they're with a highly effective teacher, and then they're with the less effective teacher or a very low effective teacher for the survey 3. So we can see a growth that might do this. Is there a delineation --

DR. COHEN: There's no way to capture that because you don't capture the kid's achievement at a different point in time; you only have at the end of the year.

So if a couple of your kids -- let's say I'm not a great teacher and there's a great teacher over there, and a bunch of my kids transfer into that great teacher's class, I'm going to get the benefit from that. That's just the case because you don't have a measurement of a year. I think it's maybe not all that likely when kids transfer or leave your class they're all going to go to teachers with particularly high or particularly low --

MR. LeTELLIER: That's what --

DR. COHEN: It's an excellent question; it's a good insight.

MS. TOVINE: For evaluation purposes, though, it seems fair for teachers if students included for their score could be ones that were there for the majority of the year, similar to the way you do student grading, and they have to be there for both counts. I mean, it doesn't seem reasonable that a student would come in in

January or February or whatever time period, and that child is now going to count in the calculation --

DR. COHEN: That is a policy decision and I'm going to hand that over to Christy. I don't know what data we have to support it, but that's the kind of recommendation I think the committee can make.

DR. HOVANETZ: Yes, absolutely, and that goes back to Gisela's point, too, is if you do require the full year attendance and it's a semester long course, is that appropriate? And so being able to look at how we're capturing attendance, data, and information, and being able to make recommendations to the Department on how to revise that is going to be essential here.

Keep in mind as we're doing this, this is with the data we have available and that was what we talked about the last time we were together and this is the best information we have right now. You all are bringing up great points in how to capture additional information by course, by period, you know, by number of days actually in attendance in that particular

course in that school. So those are the kinds of recommendations we can make to the -- this group can make to the Department. You all have the ability to say we'd like to see it done this way, but right now this is the best available information. Nicole?

MS. MARSALA: Does it have to be the same? I mean, do we have to say -- I mean, maybe for middle school and elementary where we don't have the block scheduling; can it be two and three where they have the block scheduling it could be two or three, or does it have to be the same everywhere?

DR. HOVANETZ: Again, that's a decision that the committee is going to be able to make. You all can decide how that's done. If you want it to be that that student has to be with that particular teacher for the entire year during elementary school, again it's a decision entirely up to the committee.

MR. LeTELLIER: Is this something that can be made at some point as a business rule of some sort? In talking with the Department, you know, whether or not this could be something implemented that we can look at because you were

talking about attendance and how we would do that, where we could just come up with some simple rules that would articulate what we would expect to be used?

DR. HOVANETZ: Yes, as soon as we have the data capture process, we can implement whatever business rule this committee is interested in in putting together.

MR. COPA: We're absolutely interested in those recommendations on business rules because they help inform the process that we'll put in place.

MS. FRAKES: I just need some clarification. If a student is enrolled in both a reading class and a language arts class, does that count as more than one of the same course?

DR. HOVANETZ: At this time based on the information that we have, yes, they're considered in more than one course.

MS. FRAKES: My concern is for especially for those grades that test writing, such as 4th and 8th grade, where in reading you're going to have your core reading skills, but when you go to language arts you're going to focus on that writing and I just wanted to make sure I said that because I've heard a lot of feedback from teachers who have watched the webinar and language arts is taught quite differently when you're focusing on writing.

DR. HOVANETZ: It's a fantastic point you make. Tomorrow afternoon we have a conversation about the course code directory that we don't want to overwhelm you with now, as we're trying to facilitate the process of having you all make recommendations to the commissioner. But for the process of this evaluation, we use the information the Department had for -- course code directly. The Department recognizes that this needs to be revised or evaluated at least to determine are these appropriate courses to have for the evaluation of teachers on their reading FCAT and on the math FCAT. And we have a list, the master directory of all the courses that were included. Like, I believe there are 166 for reading and 90 for mathematics that were included for purposes of this evaluation.

This summer, this committee's responsibility when we get together next will be to talk about which courses should be the ones included or required for the FCAT statewide

evaluation. So they're bringing up fantastic points. The Department has fully recognized that's for tomorrow.

Gisela?

MS. FEILD: Let me make sure I understand the growth. Gina and I are sitting in the same English 1 9th grade class, we both happen to have the same 8th grade FCAT reading score whatever that happens to be; we're both not ELL, we're no SWD, we're not gifted, we're twins. I happen to have had some issues -- for some reason I'm also sitting in an intensive reading class or maybe Gina had parents who said I'm not putting her in intensive reading, right?

MS. TOVINE: She's qualified but she's not enrolled.

MS. FEILD: She's not enrolled. We both have scores that qualify. So we're both in English 1 with the same teacher, and I have an extra reading class with you. So when we build my expected growth, my expected growth will be higher than Gina's because I'm sitting in two.

DR. HOVANETZ: Based on the recommendations of this committee last meeting, yes, there will be additional expectations if they're enrolled

in multiple courses.

MS. FEILD: Okay, so I teach her for English.

And you, who is my reading teacher, will have different growth for each of us.

DR. HOVANETZ: Well, she's not in my class, so I'm not -- but, yes, your higher expectation I will be accountable for, as well as your teacher that you're in the class with for Gina.

And again, this is a decision the committee made last time about setting expectations and setting attribution. If we want to change how that's done, it's again something we should be talking about here and now on what that expectation should be. I would like to have some results presented first so you can understand what the expectations will take and how it's actually attributed. Jon?

DR. COHEN: One thing that folks are discussing that maybe they shouldn't get lost is we kind of covered two topics. One is how do you attribute stuff and we can figure that out. Christy says we'll look at some data and figure that out. The other thing is the data collection. Right now with the data we collect

we can't tell whether a course is intended as a one-term course or a year long course. People are concerned, well, the kid's only in my class for half the course, how can that be attributed to me? We don't know if it's half the course unless we know whether it's a full year or a half-year course, right? Gathering this data and - and the committee, I would recommend that the committee think for a couple of minutes about that issue and which data you need to support the kinds of analysis, the kinds of attribution you want because right now that's not in the state data system. Particularly it's not --

PANEL MEMBERS: (Over-speaking.)

DR. COHEN: So we could have.

DR. HOVANETZ: The only thing we don't know with the identified data set is there is not a term that says that this is a full year course or if this is a semester course. So if a student is repeating a high school course, it will look to us as if it's a full year course rather than repeating a semester course, but that's an infrequent occurrence.

DR. COHEN: We know how long kids were

enrolled.
DR. HOVANETZ: Correct.
DR. COHEN: So we don't know about block schedules. We don't have that.
DR. HOVANETZ: We don't technically have that information. We have what district and school and course number and period number and we have which survey the student was enrolled in it; and if they're enrolled in the same one identical information for two in survey 2 and survey 3, the assumption is it's a whole year course. If they're only enrolled it in one semester in survey 2 or survey 3, the assumption can be it's a semester long course, but there's no actual data that says semester or full year.
MS. BROWN: I'm confused and this will ask everybody else. At least in our scheduling system that we report, there's a cell for term -- term 1, term 2, or term 3. Term 1 is for semester, term 2 is second semester, term 3 is a year long course. So I don't understand why that data element isn't available to you.
DR. HOVANETZ: We'll get that data --
MS. BROWN: I think part of the problem is sometimes schools semester-ize courses for GPA,

for kids in graduation; and if that happens after the collection period it's possible there could be some error in that, but you're right.
MS. FEILD: But in general you should be able to know. You should clearly be able to see how that student is scheduled in that reading course.
DR. HOVANETZ: Okay. We will be sure to --
MR. MOREHOUSE: I have a question. Hillsborough County, for a term 3 full year course, students get a grade for the first semester, second semester; and those grades are based upon what?
MS. BROWN: What do you mean by --
MR. MOREHOUSE: Is that material that's on that first semester?
MS. BROWN: Correct.
MR. MOREHOUSE: Then they evaluated on that. So why didn't -- wouldn't that be a test score that's scored?
MS. BROWN: Because FCAT scores are only given once a year, so that's the reason that you only have that one option. But, I mean, that's where we're trying to attempt to get at attribution more on a semester level instead of

just because we're actually converting our term 3 courses into term 1 and 2 even in elementary because it's possible for a student to transfer between two fourth grade teachers and have one term 1 and another term 2, we can see fifty percent partial attribution even though they were enrolled in the same course for the entire year.
MS. ACOSTA: I have a question actually, maybe some of you can address this as well. I think you partly answered it, Anna. I was curious how many kids are actually impacted by having semester courses for FCAT, for things to get tested on the FCAT because at our school it would be a rare student who would be enrolled in a year-long course that had an FCAT test. Do you see what I'm saying?
How many students are we talking about? Is this very common that you only have a semester course that's going to be tested on the FCAT?
MS. FEILD: I think part of the case in Miami-Dade is, of course, recovery. They're even talking now in terms of algebra making it a semester course for course recovery, taking the DOC in December. So it's really -- sometimes

kids pass the first semester but fail the second semester, so they're course recovery-ing at the beginning of the next year just the second semester or vice versa. So there's thousands in Miami-Dade, thousands who fall in that criteria.
MS. BROWN: And what we see is that is definitely the situation in our district, and there are many, many students; but when you also consider mobility which I know we're capturing, but when students transfer school to school even in elementary, when you're highly mobile you show up in multiple places but it's very important because you're still one FCAT score, and so which teachers are given attribution for that effort? It's a lot of students.
MR. TOMEI: I just want to make an observation; I think this is an extremely important conversation that we're having right now for a couple of reasons. First of all, over the course of this committee we're going to get past the point where we're only talking about FCAT scores as a measure, so Lawrence's comment is an important one to keep in mind. What I'm very concerned about is that the ultimate model that gets put in place is going to be used for

multiple purposes. It will be used to evaluate teacher effectiveness; it will be used to evaluate school and possibly district level effectiveness; and it will be used to evaluate teacher preparation program efficacy. What may work at some levels -- and I will go back to the class size -- the question I asked about class size and how stable that is between the measures. When you get a large enough end, if you're using data to evaluate a teacher preparation program that puts out a thousand students a year, it's probably not an issue. But for one individual teacher it could be a really big deal if that statistic is unstable. So we've got to keep in mind to protect the transparency and the integrity of the model for individual teacher accountability. I think that's the lowest common denominator and I think that has to be an important outcome of this committee.

So these conversations and comments like things are infrequent, well, if you're that one teacher that experienced that infrequent event, that doesn't mean a lot to you. We need to protect that teacher as well.

So again, I just want to emphasize that I think this is an extremely important conversation we're having.

MR. LeTELLIER: That goes back to the last time we met the reason why we took the norm referenced model and threw it out the window, the quantitative -- I think it was called the quantitative based because we were concerned about that lower two percent of teachers that no matter how good they were doing, they were in that lowest two percent, so you don't have a chance of getting support as effective, and I think that that is something that I will agree with that that we can't take the risk of just saying, well, we may have three or four teachers, I'll just put out a handful, that they may go by the wayside because that's certainly not fair to the teacher and it doesn't speak to the integrity of how we evaluate people.

So I agree with that and I think that the end result that I would see is moving towards a system that would go with the individual teacher for what they are doing so they're accountable for their work, and I think -- you know, in Florida we're all talking about aligning

everything to business models. You know, accountability is kind of something we haven't talked about too much, but all this stems out of a business model of being accountable for what your job is and making sure that you're reaching those goals or exceeding them. Well, if we're truly going to do that, it would have to be something that that teacher is able to reach. So I just wanted to piggyback off of that.

MR. TOMEI: I think part of that overall goal has to be that we need to do everything we can within the legal limits that we have in terms of what variables we can put in the model to mitigate to the greatest extent possible the unintended consequences of discouraging the best teachers from going into the locations that are in most need.

MR. LeTELLIER: Absolutely, and I just thought of this about an hour ago, is we had talked about unintended consequences. I can't remember what you -- what's the word you used last time?

PANEL MEMBER: (Inaudible.)

MR. LeTELLIER: And I've been dying to say this and I haven't said this publicly, but I

think we should: Every profession is going to have people that try to play the game, that are going to try and skirt around things and try and make it work for their benefit, but I would like to think that the majority and not just a simple majority, but the vast majority of teachers are not going to try to game the system. So what I guess we don't want is the reverse where we make it so that, okay, we've taken away the perverse incentive but we've made it so that it's not sustainable or obtainable to the teacher that is trying to do their best. I think that's something that's been weighing on my mind for the past month is making sure that we don't just rule out things so we say, well, there's a perverse incentive there for a handful of people that are going to take advantage of that and maybe there's some way that we can address that rather than just throwing out a specific variable that we figure, okay, how can we address if that happens.

But I really -- you know, being a teacher for the years that I have been and the people in this room, the teacher, the superintendents, I mean, the parents here, you guys know the majority of teachers care, we love what we do. We want to help kids and the last thing on our minds is how do we game the system. So I think that's very important, something that I just want to share.

MS. EDGECOMB: In this discussion and maybe I'm not hearing it correctly and maybe I shouldn't get bent out of shape about it. I'm worried about uniformity I'm hearing. Is that something -- because I hear how people are capturing and coding and talking about courses and when they end, how they end, and what they are and what they are not. If we're using the motto that's supposed to be fair and formalized instruction to capture and save information, isn't there some importance and value in some uniformity here?

DR. HOVANETZ: There is. I absolutely agree. The Department has a directory that defines each of the particular variables that we are using, so I think they are well defined. It's assuring and again they have been reporting these variables to the Department for in many cases decades. In the same way with actual data elements, definitions, and particulars on how

they're supposed to report it and when they're supposed to report it, each one of them is very defined.

It's assuring that they're following those proper procedures, which I'm assuming the majority of the districts are but continuing to reiterate as soon as we do determine which are the control variables or the covariates we're going to be including in these models to highlight those when they do that -- and I ask for -- when they're at the consortium meetings to be able to highlight and say these are the particular covariates that are included in the models, these are the data definitions that you need to be paying attention to, and then also through the review process to pay particular attention to these covariates, but I agree, uniformity is going to be paramount for insuring the accuracy of these calculations. I think that there are already a lot of processes that the Department has in place to insure that and continue to use this data and information that's going to improve the quality of it, too. But I absolutely hear that that is a concern and I think it's going to just take additional PD and

technical assistance with MIS directors and others who are reporting on that information up to the state to be sure that they are following what's in the data element dictionaries.

MS. BROWN: If I can just piggyback on that because uniformity is so important to the validity of the entire process, and in my experience what I can see the writing on the wall where we're going to need to go is how we clearly define those business rules for eligibility because in some cases we may not have uniformity in say, for example, course assignment. But we can create uniformity by creating the right business rule that looks at those in an either/or situation, but then also the process itself starts to find the anomaly and when we start to find the thing that stands out as different then there's just that oversight that starts to correct everything towards uniformity.

We've discovered that in our own district when we're looking across large numbers of schools and how individuals are scheduled, et cetera, and we would do that major oversight and you think that you have every one common, and

then when you start to look at the data set you go, what the heck is this? And then that is where you go to that situation and say you've got to fix it. And so through the process we'll get there, but I can so respect and I would like to piggyback on that, that that uniformity has got to be part of the goal in order for there to be any kind of validity for the comparisons that are going to be made.

DR. HOVANETZ: And you make a fantastic point, and when this committee makes a decision about which model is actually going to be used and the commissioner approves it, the first AIR responsibility is to generate those value-added scores for teachers in Florida and have those posted this summer. That would be our first opportunity to start looking for any type of data anomaly for districts and schools to review, to start looking at what the data and information looks like in order to identify the anomaly where we see something happening in one of the districts but not the other 66 and be able to start identifying where we need to provide additional technical assistance.

MS. NOYA: Also, if you leave up to the

interpretation of every district then we start doing everything differently, and that's been one of the problems across the state for years. So transparency, teachers are asking -- when in the summer? Are you going to do this late summer so when they come back and they're in a rush, they won't even see the transparency?

DR. HOVANETZ: Juan is panicking because of what I said, but there's --

MR. COPA: I just heard an audible groan with the word "posted", but this is very -- districts have been very interested, of course, as we're developing part of Race to the Top their evaluation systems, information needed to inform those decisions regarding their evaluation systems, and we are committed to providing districts with data on the model that is eventually selected to help them form decisions. So it's really a provision of data to districts in a useable form to help folks start to understand what it means, make informed decisions about how to apply an evaluation system. So that's the key point we'll accomplish this summer.

MS. TOVINE: Once those results are

actually run and they go back, some of them are selected, will this committee come back together, say, in the fall or whatever time period that is to kind of see how that's going and maybe look for some of those problematic areas and maybe could make some sort of recommendation how to do it differently?

MR. COPA: Gina, you didn't realize this was a four year commitment? Very clear, this is definitely not the last meeting of this group. This group will continue to meet throughout the four year process, not only dealing with -- we have this -- there's the sense that this is the end for the FCAT model as Cathy said at the beginning; no, this is really the beginning and so we will continue to review that. We will start working on those other models, the EOCs that will eventually come online starting with Algebra 1 which was just administered last week, so yeah, there's a lot of work to be done over the next three years.

MR. LeTELLIER: Juan, I think if I understand you correctly and what I think will make everybody feel a little better is what we're looking at is presenting the data with the

model that we use so that people can just look and say, okay, here's where we're at. We tweak it from there, nobody's jobs on the line because of the data coming out in the summer. I think that's what people are kind of --

MR. COPA: Absolutely, yeah. This is for information purposes.

DR. DORAN: That's a good point. This gives us the basis by which we can look at some of the data, changing some of those business rules wouldn't switch, for example, which value-added model necessarily look different, but it would change some of the things about attribution. So it's still us; we're still on safe ground on how we discuss what we intend to, even when the business rules perhaps don't change.

MS. FEILD: Which FCAT scores are you using? The old scale, the ten year scale?

PANEL MEMBERS: Over speaking.

DR. COPA: Correct, for the purposes of this summer you're well aware it's equated exactly to the old scale, but it's part of their work as well. It's no secret we're moving to new standards beginning this fall. It will be

applied starting with the spring 2012 assessment, so any change in the assessment will likely result in further refinement of the models as well. So again, on to this continued responsibility of this committee.

DR. DORAN: The models that we've implemented are scaled independent, so even should the scale change the models can still be applied.

MS. FEILD: No, I understand that but we're making decisions on the model based on the old FCAT, and it's possible that with the new standard and the new scores the model we pick now may not be the best model with the new data. That's why I was asking and I think you answered it by saying that we will be able to tweak the model. What I suggest is that when we have the new FCAT data and we almost have to wait two years to have two years by this hearing, we have to revisit the model because maybe at that point some of the variables we're using are not appropriate or we have to change the model.

MR. COPA: Possibly, yes.

DR. HOVANETZ: Absolutely. Four year commitment.

MR. COPA: These are all empirical questions. The commissioner must make a selection by June 1st. When that becomes part of State Board rules, I mean, that is something that would happen in the future going forward after we've had the work of this committee to look at years of performance, refinements, and models; and even when it becomes part of the State Board rule you could further refine that going forward, as well. So that's a key thing to keep in mind. It's not something that we will be etched in stone that cannot be changed. The anticipation here, of course, is this will be an evolving product.

MR. MOREHOUSE: I'm just curious. Is it possible to develop a description, a course description? There seem to be so many variations in terms of how a course is delivered, some courses will be graded one semester, some are year long courses, some are team taught, some are not taught; is it possible to have a description of those options for the committee so that any rules that we have to make in my mind, it's hard to depend upon how those courses are delivered and how they're graded and

when they're delivered.

DR. HOVANETZ: Yes, and I think we can fold this into our course code discussion that we'll have tomorrow, too, as part of the ongoing work of the committee as well. There are well defined course descriptions that have standards associated with them as well; the delivery method of those is more the district discretion on how the content is delivered, but we will definitely role that into part of the conversation.

Doretha?

MS. EDGECOMB: I apologize if this is a stupid question up front, so if it's stupid I apologize.

Is there any value, and I think I heard Gisela say this, to recommend a model that had limited number -- this is progressive -- limited number of covariates and add onto it and over time have a lot of them and take away from those covariates?

DR. DORAN: I'll answer the question. There is a criteria that we will evaluate today that we call parsimony. Parsimony looks at whether not --

MS. EDGECOMB: Oh, is that what that means? How many other people knew what that meant? Okay.

DR. DORAN: Doretha, it's not a stupid question. It's exactly a right on question. Is there value in adding additional covariates over time?

Well, given the covariates that we have we can evaluate is it good to have some or more, and we can provide you data that suggests whether or not it's good to have some, only a few, or whether it's good to have more and we're going to look at exactly that question when we go through the parsimony issue today.

MR. LeTELLIER: Can you define 'parsimony'?

DR. DORAN: Remember, when you go down the criteria, if you have a question, we'll define everything, we'll talk about what we're looking for and why you should care about each one of those things.

DR. COHEN: Other things being equal, simpler is better.

MS. EDGECOMB: Absolutely.

MS. HEBDA: Okay. To keep things simple, we'll take a 15 minute break. We'll regroup at

20 minutes till.

(Whereupon, a short recess was had.)

DR. DORAN: The part of the conversation that we're about to go in right now is very data centric. We're going to start looking at some of the data relatively soon. We've got a couple of slides to get through. The conversation we had this morning really should underscore how important the quality of the input data is into the statistical model. As I walked outside to go to the Einstein, which was closed to get a little snack, I looked at the guy laying the bricks and he's being so very careful to make sure that each and every brick fits right into place, and if it doesn't he pulls out, he's scraping the bricks, making sure they fit into place because in the end that thing has to look perfect. The reliability and the validity of these models rests tremendously on the inputs that go into this. Now that was an extremely important conversation and we'll continue that conversation; at this point we're going to transition a little bit in from some of those rules and decisions that were made at the last meeting, which were fully amended by the way.

There's no reason you have to hang your hat on the things that were cited, as far as I know, in terms of going forward when this becomes operational.

But we're going to look now at some of the data, and we will evaluate some of the models based on the different categories that we've laid out for you. So let's talk about a couple of things and a couple of terms.

One, let's talk about this thing called a deviation from an expectation. Given prior scores and other characteristics of kids, whether or not they're ELL, special ed, gifted, so on and so forth, enrolled in two courses, we have what is the average score of similar students. That's what's called an expectation. If you recall back to that scatter plot where we had those regression lines, the expectation is that line. Remember that line changes according to students. So with students you have all things being equal similar prior scores would have the same expectation. Students who score above that or below that have a deviation from that expectation. In statistical terms that's referred to a residual. It's the deviation from

that expectation.

What score did the student actually get in testing that's sometimes called an observed score. It's the score that the student actually received. So we have an expectation, we have an observed score, and this is the deviation that gets aggregated, not residual. Essentially, it's aggregated and it's a little bit more than a mean but it's equivalent to a mean. We take those residuals and we aggregate them within a teacher's class. We formed some decisions about whether the teacher had high value added or low value added.

Expected growth is a little deviation from the expectation. Later on in the presentation, we'll show some slides about differences in expected growth on different sub-groups of students, so let me define what expected growth is. We have the expected score, like I talked about here, which is the predicted score -- given your prior score and any other characteristics about you, what is your predicted score -- minus your prior score, the first lag and the most recent lag. We call that for each kid the expected growth. We have a

predicted score and we have a prior score, if you just subtract the prior score from that predicted score for each kid you get an expected growth. And that's a statistic that we formed some summary statistics on and we will show you later on in the presentation.

Christy already went over those.

Let me talk briefly about this. The expected scores change in terms of their definition depending on which model we look at. Now, for each model we form an expectation, and I'm going to put an extra word in there. We form what is called a conditional expectation. It's conditional on what your prior score was. Kids with two different prior scores have different expectations, but kids who have the same prior score have the same expectation. It's also conditional on whether or not you were SWD, special ed or gifted or so forth. So, for example, I think Gisela was asking this question earlier, two kids who are exactly identical on all of their characteristics -- prior score and categorization in gifted, special ed, and so forth -- they're the same on all of those. They have the exact same predicted score, everything

being equal, but you change one of those things.

So suppose students are exactly the same in terms of their gifted coding and SWD coding and everything, but they have two different prior scores, they would have different expectations and whether or not one student had a higher or lower and depending on what their prior score was. So all things being equal, kids with similar prior scores have similar predicted scores and similar growth expectations, but when those things change they have different expectations.

DR. COHEN: Harold, the numbering on this chart is wrong.

DR. DORAN: In fact, you know what --

DR. COHEN: We have talked about this; this is going to cause confusion.

DR. DORAN: Why don't I skip this because this really doesn't add substance to some of the criteria we're going to go through. We could talk about those things contextually as we encounter those models. Thanks, Jon.

All right, let's provide a big picture before we start delving into some of the specifics. Here's a big picture. There are a lot of models and many different variables. We have eight different models, seven grades in two subjects. There are more than 100 models that were estimated, and we have 18 different criteria by which we evaluate the models. If we wanted to spend time looking at each model for each grade for each subject against each criteria, we would need to be here for four years. This wouldn't be a four year contract; we'd be here for a very long time. It's not viable. Just simply cannot do it.

Now we're going to try and consolidate this information based on some things that seemed reasonable to do; I'm going to show you how we try to narrow some -- how we're facilitating this conversation today. The point here is that there is a lot to look at there are many different variants that were estimated from many different grades and subjects. But as we're going to show you, the key results are consistence across all grades. If we saw that models behaved differently across grades and across subjects, we would need to have pulled those models out and examined them. Why does grade seven Model 1 look so different than Model

1 in grade four in math? We don't see the behavior of the models that would have required that. I'm going to show you in a moment.

Essentially, what we're going to do is we're going to choose a particular grade, grade seven, we chose grade seven because it's in the middle and because it's like the other models similar in its behavior to all the other models in all the other grades. Let's show you why.

Sorry, just a couple more things. Housekeeping. Actually, Jon, I'm going to turn this over to you real quickly because I know you have some of these statistics in your head, and before we look at some of the models let me talk through this and then I'll give the mic to Jon.

One, the estimators we're using for models are unbiased, consistent, and yield accurate standard errors. Let me be clear about one thing: The models that we estimated, that were developed, we have two kinds of data, real world data -- the FCAT data for kids who were actually tested in the state of Florida. The results that we'll present to you today were based on the real world data, but when we do development of statistical models and software, we do

extensive testing to make sure that the models behave in the way they're supposed to behave. We look at the standard errors which we'll define in just a little bit. We want to make sure that the models and the software behave in the right way.

There's an extensive amount of testing that we do on simulated data. Simulated data means we make up data according to some assumptions and we test out our models and we make sure the models give us back answers that we know they're supposed to give us back. Jon's going to expand on that as soon as I give him the microphone because we did an extensive number of simulations to make sure that things were working as expected. You want to talk about that?

DR. COHEN: Yeah, I'll talk briefly about it. There will be a technical report that has all this in here, but we're going to -- when you estimate a model you want to make sure that the model is unbiased, right? We want to have unbiased estimates of teacher effects and we want to make sure that the way we're doing things statistically isn't introducing a bias.

What is a bias? If there's a true value then any deviation from that true value is a bias, right?

So in the real world we don't know how much value each teacher has. You can't know that about teachers. The teacher fairy can't land on your shoulder and whisper in your ear "Miss Jones is absent, add 10 points of value". But what we can do is we can make up data, right, and we can make up the data in different ways. So we know the true values. So what we do, we define a process for making up the data and then process the random ones. You choose random trials from this population. We generate 200 different data sets and then estimate the model 200 different times. Again, this is on data we have made up so we know what the true values are.

Is there a pen here? I'm always happier when I have -- look between, oh, yeah. Let me roll this out here. Okay.

So every teacher effect that we're going to estimate under any model, this is a value-added scale, and this is -- you think of it sort of as a probability. It's a density function of the value-added, we'll just give it a little Greek symbol there for the value-added just to have something to put up there.

All right. So let's say a teacher's score is right in the center. So your value-added score is -- I don't know what the number is, maybe you're adding 20 points. That 20 points is a point estimate, but it's not really a point estimate because we don't know. All we have is we a probability distribution and they tend to come out to be normal probability distributions -- there are proofs that statisticians like to think about. So here's your point estimate. Each point estimate comes with a standard error, right? So you know that the teacher is very likely to be at this point, a little less likely to be at this point, way less likely to be down there, really unlikely to be up here. But it's a probability distribution; it's possible.

So we estimate these standard errors. When we talk about precision, we're going to be leaning on these standard errors. If the standard error is smaller, the model is more precise. I have less uncertainty. If I had a curve like this, I would know with some

certainty that this teacher was in at least in this range and did not fall out here. Is that clear?

So more precise is better. It sounds better, it is better.

So you can't observe the standard error, so there are statistical formula that you use and for these models they become very complicated. So we tested the statistical formulas for the standard error by generating lots of data where we knew the real answer, and we tested -- we got the standard errors and we counted the number of times that the estimates fell out in this range. We counted the number of times the estimates fell out in that range and the same on the other side, because the standard error and the normal distribution tell us we ought to have five percent following here plus here. We ought to have ten percent following here plus here. For all the models that we're showing you, we wound up with unbiased standard errors. So the standard errors did -- when you make up the data and you look at it over and over, it does reflect -- the standard errors do capture and tell you the right percentage of teachers who

would fall outside that range of the true value.

So it's important, it's technical. There will be a technical report on it for anyone who wants to read that. A couple of people here indicated they would like to see that. We can show you results if you want to see them. They look like numbers. We expect 10%, 10.1%, we expect 5%, we're looking at 4.89%. So we're getting lots of data back for lots of different models on lots of different situations. No one ever tells me they can't hear me.

So I think that's it. The models are also unbiased. We got on average the right estimates back and consistent, that means as the sample size gets bigger the estimates get more precise and all the models we're working with have that characteristic as well. So as we move forward, rest assured standard error we're talking about are accurate standard errors when the data corresponds.

DR. DORAN: All right.

MS. BROWN: Before you leave, when you talked about precision, could you kind of define for me precision as accuracy?

DR. DORAN: Would you mind if we hold --

DR. COHEN: It's easy. This standard error curve is more narrow, it's more precise. This means the estimate is more precise. If it's wider, it's less precise.

MS. BROWN: Did you mention something about accuracy, though? I thought -

DR. COHEN: All the estimates are giving us accurate estimates, unbiased. So we're taking accuracy and we're breaking it into two pieces, unbiased and precise. Is it centered around the right number and how much variation is there?

DR. DORAN: We're going to talk a lot about precision in just a minute here. So here's the roadmap at the bottom of the slide. There's actually a little bit more. This refers back to the crib sheet that we've given you. We're going to look at the model's different size, the effects attributed to teachers. We'll talk about that, how precise the estimates in terms of what they yield, what are the expectations of growth established for different groups of kids, and what is the impact of various models on different groups of teachers?

So this is big picture. Now we're going to start looking at some of the models across all grades in both subjects, and this is going to be the first thing that we're going to show you is a line graph that shows the magnitude of the effects. Let me actually just show you what we've got here. We're going to start with reading. This is all models, all grades. This slide and the next slide are the only time today that we will show you a statistic on all models for all grades as we go forward and as I'll explain in this particular slide here, this slide provides the justification for why we only look at one particular grade but all models. We always look at all models, but only for a given grade. In this particular slide, we look at all models for all grades, okay. This is the last time we'll do this, the only time we'll do it.

All right. Here on the X access we have each grade and on the Y access we have what we're calling the size of the effects. Now this here, the model, remember looking at your sheet differ in terms of a few characteristics. Some of them only have teacher effects and some of them have both the teacher and school effects. We'll actually define what that means to have both teacher and school effects in just a little

bit.

But what we see here is that these are the covariate adjustment models and they behave very similarly. The effects are comparable across all grades. We don't see the lines criss-crossing in very unpredictable strange ways. We don't see something that looks extremely anomalous suggesting that Model 1 in grade four is behaving very differently than Model 1 in a particular grade. There are very tiny differences in the models across grades. So, for example, in this particular model here which is, I think, on this square Model 1 it's behaving similarly across all of these grades. The effects always appear to be the largest. We'll see why in a little bit. Then in here you don't see criss-crossing.

Up here is the simple differences model. Now it's measuring something different as we explained before, all right? It's not measuring the exact same thing as the other covariate adjustment models are, and so we see a big difference here between its effect and its relationship with the other models, all right?

What we've done over here on the right,

we're going to introduce this now but we're going to look at this in a very detailed way in just a little bit is what happens when we add the school effects back into the teacher effects. In part, we're not ready to evaluate this criteria yet because we really haven't defined why you should care about school and/or teacher effects just yet. This is a high level overview.

What happens if you add those school effects back in; essentially what we see is that the models behave similarly again. So adding school effects back in as opposed to having only teacher effects in some of the models causes for the behavior of the models to be similar. Let me go to the next slide because we see something similar in math, and I'm going to revisit the key point that we're looking for here. The key point, the key take away in the slide that I showed you before, in the slide that I'm showing you now is that the behavior of the models across grades for both reading and math is comparable. They're similar. So we're using this as a justification for why later we're only going to look at a single grade all models.

You're not ready yet to evaluate any of the models based on the criteria, the precision, the accuracy, and so forth just yet. This is just the high level overview.

We see the same thing in math. We don't see the model here and here and here and criss-crossing lines which would suggest that Model 1 behaves very differently depending on what grade you're looking at. That's not what we see. We see a consistent pattern here and when we add the school effects back into the teacher effects. It's the only inference that we really want you to draw from these slides. That is, models behave similarly across grades.

MR. LeTELLIER: Can I ask a question? How large of a sample group do you look at because I don't think we have that anywhere in the info?

DR. DORAN: It was -- I'll tell you, I don't know the exact numbers but we do this by grades. The number of teachers in each grade ranges from 7,000 to 10,000 or so teachers per grade --

DR. COHEN: It's all for the entire state.

DR. DORAN: Yeah, the entire state. There's no sample, it's a population of kids.

MR. LeTELLIER: And the second thing is when we're looking at the effects of the variables when deciding this, I know now we're looking at sort of a macrocosm sort of dealing with the big picture; are we going to look at how those variables may affect an individual class?

DR. DORAN: Yes. Not a particular teacher, but on average across many teachers, yes. We see exactly that.

MR. LeTELLIER: Versus just plugging in to the whole state is what I'm saying because if you -- in other words, you may have with the attendance issue of a kid that missed 60 days out of the 180 days, that certainly hopefully is not the norm, but that may have great weight for that teacher in that class. So if we only plug that into those 10,000 teachers, it's going to be a blip on the screen and won't show up. That's my question.

DR. DORAN: All right. We don't look at any one teacher where there's any particular impact, so does adding that variable in change your ranking from high or low. But one of the things that we do look at is how correlated,

what's the relationship of the teacher effects across the different models. That tells us whether there's a flip-flop of teachers being highly classified and under one model then maybe being classified differently under another model. We do look at that. I can tell you now even though that's something we'll talk about a little bit later on, that the consistency of the classifications across models is virtually perfectly correlated. Okay?

MR. COHEN: Harold, we can say something -- when we get into covariates we can talk a little bit about what the likely effect of differences in rates of attendance would have on individual teachers' value-added lists. We can look at that. We can talk about it. We didn't pull out any individual teacher. There's data as to that question and again we're talking about covariates.

MR. LeTELLIER: I wrote it down. I get the general thing, but I don't know if anybody else is understanding where I'm coming from is that not missing something because it's in such a large set of data and making sure that it -- in other words, we make sure that a variable can

affect a teacher in a certain way and we're going to take that into account. Now obviously if it only affects one teacher across the state, is that the norm? No. But what we're saying is if all teachers that would have a kid that missed 60 days, let's say, whatever teacher that was, is that an effect? Are we looking at that in that way?

DR. DORAN: Yes. Before we go into the different effects including attendance has on the predictions of the --

MR. LeTELLIER: Attendance was just an example --

DR. DORAN: Of the different covariates.

MR. LeTELLIER: Okay. Thank you.

DR. COHEN: I think it's math per grade says a kid who missed 60 days, everything else being equal, would have an expected scale score of seven points less. So you would have one kid who has an expectation of seven points less among your whole class. So it may matter and attendance is something that you want to think about. You can't learn if you're not in school, right?

DR. DORAN: One thing that I didn't tell

you at the beginning of the day is for grade seven. We brought data with us, some data files, so if there are things that you're curious about that we don't present or there are some questions that we can look at, if I'm talking Jon will crank through and maybe data analysis done or if somebody else is talking, I can crank through an analysis and we can try and answer some of those questions. There are no instances where we look at one teacher and look at whether or not that particular teacher is changing. We look at the population and how things are behaving across the state.

MS. WESTPHAL: When you're looking at the models, and I understand the purpose of this was to justify why we just picked seventh grade to look at, I see it in the reading and I see it on that one. Is that not a significant difference on the left between sixth grade -- fourth, fifth, sixth grade versus seventh, eighth, ninth, tenth?

DR. DORAN: So there's an interesting finding here, all right. If I'm following what you're saying, the significant difference is that we see this downward trend.

MS. WESTPHAL: Well, no, the consistency of the models seven, eight, nine, ten versus four, five, and six.

DR. COHEN: No, I think the important thing to take away from this is that the model -- even though it's developmental scales, the scales are kind of different between grades, and so you don't necessarily expect the same numbers in each grade. What's important is the relative ranking of the models. Model 1 always sees the biggest teacher effect, Model 2 always sees the next biggest effect, so the lines are parallel. That's what you're looking for in this. There are differences in the estimated effect size across grades and there's a difference in the pattern between math and reading. But that confounded with the differences in the measurement itself. We have run everything and we've looked at all the data and the big findings stay the same. We just didn't want to try and present, you know, two subjects times seven grades for this many models. You'd never be able to look at the data.

MS. WESTPHAL: Well, let's say we're looking at grade six. Two of the models or

however many lines, it's hard to tell color-wise, but there seems to be a bigger gap between those two models than there does once you get to 7th, 8th, 9th, and 10th.

DR. DORAN: You're right. There does seem to be a bigger gap. Let's look at the issue here, though. This particular model is harder to observe because the lines are virtually on top of each other. This particular model in grade six shows the largest effects and what we see as we go down, even though the gap dissipates that is still the model that shows the largest effects. So the fact that there is a virtual difference here is not necessarily meaningful. If we saw that this had the largest effects in this particular grade and the smallest effects in this particular grade, that would be substantively interesting and would require us to pull that model out and present it to you and say this model behaves differently across different grades, we need to look at it.

If the relative ranking of probable models gives teacher effects across the different grades, it remains almost constant.

MS. WESTPHAL: That makes sense.

MS. BROWN: I have a question about -- and we may get there and it's okay if you tell me we're going to get there. But when we're looking at this, it's obvious we're going to have different patterns subject-wise, but ultimately we have to pick one model. So are we going to have some analysis about how one model reacts in subject one and subject two? In other words, will it have the highest effects in both subjects or high here, low here? We need to make that decision.

DR. DORAN: Yes, essentially what we're going to do is we're going to look at all the models in one grade and you're going to evaluate that in reading, see how it behaves in reading and how it behaves in math. I'll tell you now that the models behave similarly across the different subjects.

DR. COHEN: In most regards.

DR. DORAN: Pardon?

DR. COHEN: In most regards.

DR. DORAN: In most regards, yes. So you'll see, I think -- I don't know, but part of where you're going is you're wondering might I choose Model 1 for reading and Model 3 for math

or something like that?

MS. BROWN: Or if we are forced to make one choice, we have to make a decision; so that's important.

MR. COPA: You're not constrained in that regard. If one model fits reading better than math, you can have a separate model for math, separate for reading.

DR. DORAN: So let's actually start looking at some of these results and some of the data. We're going to structure the criteria around the following four issues. We're going to always have a question, what are we looking for? What do we want to know when we talk about precision? Then we're going to give you a statistic. What statistic gives us some evidence. Then the third thing is what are we talking about in that statistic? How do we know? What are we using to judge it by? Then last why do you care? Just sort of a simple this is why it matters to you.

Then we're going to show you data and then we're going to try to summarize what we observe in the data.

We're going to first talk about precision.

Precision in terms of its questions, what characteristics of value-added models lead to more precise estimates of the teacher effects? The statistic that we're going to look at is called the standard error. We're going to look at the standard error of the teacher effects. Let me talk about what a standard error is.

It's actually something that's very familiar in polling, so for example if you look at the president's popularity rating and the president's popularity rating is 50% plus or minus 3 percentage points. That's typically what we see, right? That means, you know, it could be -- it's a little bit more than this but it's somewhere between 47% or somewhere between 53%. We're pretty sure it's within a small range. That's standard error.

Now supposed the president's popularity is 50% plus or minus 20 points. Well, is the popularity when I say 20 points, 30, or 70? That's a big range. It's a big range of uncertainty. The standard error tells you I've got a statistic, the president's popularity is 'X' 50%, but how certain am I that it's within that range. It's a certainty statistic. Well,

if the standard error is big, we don't know. If there's a big range, it could be anywhere in there or is there a small range? I actually know it's in here. There are some statistics that we use, this is called the standard error.

A small standard of error is more desirable than a large standard error. Let me apply that to a teacher effect. We're going to get a number; that number is a teacher effect or a point estimate. That number -- and then we get the standard error that tells us, well, how certain are we that this is where the teacher's ranking really is? A small standard error tells us, the variability is pretty small. A big standard error means we've got a lot of uncertainty. It's not very precise. So what are we looking for? We're looking for a model that yields with other things being equal smaller standard errors. This is what we want to see. These are data that we're about to show you.

Why do you care? Well, a standard error tells us that the estimated teacher effect is more precise. You don't want to estimate teacher effects with a lot of uncertainty.

Essentially, what you're saying is here's the teacher effects, the teacher has high value added. How certain are you? Not so certain at all. That would be bad.

Instead, we want the inference to be, here's the teacher effect, the teacher has high value added. How certain are you? Pretty certain. Standard error is relatively small. That's our goal, that's what we're looking for in the next couple of slides.

What I'm showing you here are what are called box and whisker plots of the teacher's standard errors. We've got them ranked by the different models, model 1, 1A, 3A, 3A1, 3B, 3C, and Model 4. Refer to your little cheat sheets so you can remember which is which.

Now this black dot in the center is the median standard error, okay, the median standard error. We see in all of these covariate adjustment models, Models 1 through 3, that they are on the same scale. I'm going from about 3 to a little bit about 30. We're on the same scale. So those can be compared very easily and we can see the black dot is pretty close to being about the same, but you can see small

differences. Model 1A has smaller standard errors than Model 1 on average. That's what the black dot represents. You can look and you can see the standard errors are smaller under a couple of models, Model 3 has slightly smaller standard errors than almost all the others, right?

Now Model 4 is on a different scale. You can see it ranges from 0 and this goes up to 800. There are some outlines up there. Now it's hard to compare Model 4 to the other ones because it's on a different scale. So one of the things I'm going to tell you is the average standard error of the simple differences model is much, much larger than the average standard error of some of the other models. Now that's what this little black dot is.

What we also see is a distribution. This is the standard error of the 25th percentile and this is the standard error of the 75th percentile, and then this shows the standard error of the 5th percentile and the standard error of the 95th percentile. So what are we looking for in these box and whisker plots? What would be desirable? Well, we want to see that the black dot would be over to here to the left indicating that on average it has a smaller standard error -- by the way, this is math, grade seven math. Two of the black dots over here to the left indicating that the model has a smaller standard of error relative to the standard error of other models, these are standard errors of the teacher effects.

Then we want to see smaller spread. We don't want this box to be big. That's not desirable. That means on average of small standard error but there's a lot of variability. What we want is the black dot to be on the left and that box to be smaller. That would be desirable. So what do you observe? Any reactions to the teacher effects standard errors?

MR. LeTELLIER: Model 3C, 3B, and 3A all are about the same, the spread and they all have similar low error range. That's what I would quickly say.

DR. DORAN: That's a good observation. I want to go to another person, but I want you -- I'm giving you an assignment. I want you to go to your cheat sheet and I want you to find any

characteristics of those models that they have in common.

Any other observations about box spots?

MR. TOMEI: They all have two lags, they all have school effect and teacher effect.

DR. DORAN: An interesting observation. Let me go over to reading. Go to reading, okay? We're looking at the same charts, box and whisker plots, reading grade seven, and we want the inference to be the same. That little black dot we want it to be to the left. Now, remember, this one is on a different scale so you know it looks like it's on the left. It's on a very different scale. Its standard areas are larger. Over here you can compare the black dots. You can see that this one is smaller than this one on average and this one is smaller than this one, and we want less spread.

Can somebody make an observation here?

MR. TOMEI: Same three models. 3C has the smallest of those.

DR. DORAN: Right. So one of the questions that was asked earlier today is how will you remove some of the subjectivity when we're making these rankings? Well, these are data,

these are empirical data. We're looking at criterion, standard errors. That is an important statistical criterion. We're not bringing to you any opinions about which model we think works differently here. We're showing you the results of the empirical evaluation and we can tell you now that these models behave similarly across grade attendance to the similar trends we see about these models in the different grades, but notice what you see here. I was asking the question the behaviors of the models across the different subjects. We see some relative consistency on this particular criterion. It's an important criterion, precision is important but it's not the only thing. There are other things that we're going to look at today.

Let's do something real quickly here. We can just tell you an observation that you've already made. Suppose we were to do the ranking, which one took the smallest standard errors on average, and if we rank them what are the characteristics of those particular models? Here's a chart that summarizes this for you. So Model 3C has on average in both subjects a

smaller standard error. There are small differences between the different models. You see 3C and 3B are consistent in the first two rows and both subjects, but there's a little bit of a difference. But you don't see that that 1A is over here and here and it's minor differences, but pretty close.

So at this point one of the things that you can do in your little notes make some decisions -- and we're going to give you just a minute here -- about what do you like about precision, in that column precision, now you can make a judgment about the terms of their precision. I'm going to give you two minutes to do whatever you want on that page plus ten, a one, or a plus or a minus. We're going to do this as we go forward with today because by the end when we get to the last criteria, we don't want you to be overwhelmed with "I forgot which model was good at precision, which model was bad" --

MR. LeTELLIER: We are looking at reading and math for all these categories?

DR. DORAN: Yes, for all these categories.

MR. LeTELLIER: So I'm going to suggest just because -- I think it was Pam or Anna had mentioned about if we're bound to use the same model for both and Juan said no; maybe we should put a slash here so that we can look at both.

DR. DORAN: Yeah, put a little line through it.

MR. LeTELLIER: This way we're evaluating for reading and math, and at the end of the day it's the same, great, but it'll be easier to remember, I think.

DR. DORAN: Yeah, I wish I had given you two sheets.

MS. FEILD: Can I ask -- when you ran them going back to the models with the two year lag ones that seemed to have better precision, did that include every 7th grader in the state regardless of whether they had two data points or three data points, or was it only --

DR. COHEN: No, this only included the ones that had both data points.

DR. DORAN: Let's say something about this.

You only put a student for two data points --

MS. FEILD: No, no, we have -- these models -- the top models all yielded two lags. That means they would have had three data points.

DR. DORAN: Correct. The correct score for the --

MS. FEILD: In other words, your fourth grade model for that component would never have two lags.

DR. COHEN: Yes, we didn't estimate it for the fourth grade for exactly that reason.

MR. COPA: That's a great question and it leads to the policy implications of you need to -- you know, considering precision but also what are the consequences of selecting a model that requires two lags.

DR. COHEN: Right.

MS. FEILD: Right. That's why I'm asking -- but the first question is what was plotted before.

DR. COHEN: It included only students for whom you had both data points, and there's nothing preventing me from making a recommendation that says use two data points if you have it; if you don't, don't.

MS. FEILD: Okay. So if we were to choose a model that looks at two lags -- 3A, 3B, or 3C -- let's say we all decided on 3B right now, whatever. Does that mean that we would only

generate -- for a teacher we would only generate data for the kids who have two lags?

DR. DORAN: No, no. We can make a decision so that when students have two prior test scores, there can be a decision, a policy decision. When students have the two available test scores, you use them and if a student has perhaps only one of the two, then what we would do is we could put in what's called just a code that would indicate that one of those two scores is missing and only use one of the scores. Of course, if the student doesn't have anything, then you can't use them at all. But I don't think we can use them -- but you can make a decision.

Use two where available and when it's not available use one of the two, or maybe you just use the most recent of the two.

MS. FEILD: So if we were to use that combination meaning however many it's one set. If you have two, three years with current, you use three. Do we know how that affects the precision of the model?

DR. COHEN: Not -- there's a standard error for each individual teacher, each teacher's

estimate. So if a teacher had a lot of students with only one year of prior data, the standard error around their estimate would be a little bit larger.

MR. LeTELLIER: So that wouldn't be as precise.

DR. COHEN: Yes, we have a little less precise estimates for teachers and a lot of the kids in the classroom who didn't have the extra data. You have more precise estimates for teachers who did have the extra data.

MR. LeTELLIER: So what would be the teacher implications on that one.

DR. COHEN: Well, that depends on your classification which we're going to get to at the end of the day. But you have variability in teacher classification that are much more important than that right now. The teacher teaches a lot of students; they're going to have a much more precise estimate than if they teach a few students, and there are big differences in teacher precision. If we page back a couple of these, let's see -- you mentioned Model 3B, it's right here. You have some teachers -- 25% of your teachers have standard errors of 14

developmental scale score points or less; 25% of your teachers have, I don't know, let's call that 18 developmental scale score points or more. So there's some variation in that.
MS. FEILD: Right. I understand that. But let's say that was the model we picked and let's say that we're telling teachers this is your standard error, but really if you happen to have some kids -- happier kids don't happen to be missing one data point; your standard of error could be bigger.
DR. COHEN: And it would be reported as larger. You would know that it was larger; we would capture that in the model and you would know that they had a larger stay there, just like you know that this teacher has a standard error and the ones down here have the smaller standard error; you might have --
MS. FEILD: Right, I understand.
DR. COHEN: -- more up there, yeah.
MS. FEILD: And we know that the more years of data you have the more precise you get, but teachers have to understand that if their children do not have the multiple years there will be more error in terms of that child's --

DR. COHEN: That's right. That's right. No one asked --
MR. TOMEI: Just for clarification, if you would, go forward a couple of slides for me if you would to that rank ordering again. We've talked about precision, really two aspects of precision. One is where is the midpoint of that bell curve if you will and the other is how compressed is the bell curve? It looks like the rank order is based only on the midpoint, not about the width; is that --
DR. DORAN: That's right. So this rank order in here is based only on which one has the smallest standard errors on average. If we're to rank them on which has both the smallest and the less variability, the table would be more complex.
MR. TOMEI: And the reason I'm asking that question is I think if you did that on the right there on a map, 3A would hop above 1A. That's what the graph suggests.
DR. COHEN: Let me -- let me. It's not exactly the variability that we're interested in. If I had this and I had a big wide bar that went all the way down here, that would be good.

I'd have more teachers with smaller standard errors. I want more teachers -- so really, what you want is you don't want any - you don't want a lot of teachers falling in high ranges; you don't mind a lot of teachers falling in low ranges. So you look at Model 3, the 3A one, and Model 1, they're both pretty variable, but the 75% of the teachers in the Model 3A one have a standard error of less than 24 points, whereas with Model 1 you wind up with -- is it 1 or 1A -- in Model 1 you wind up with, I don't know, the 75th percentile and about 27 points. Right? So you want more teachers down at that end of the scale which is the reason we display it that way so you can see where you have a lot of outliers up here.
DR. DORAN: Pretty much what you see in Model 3B.
DR. COHEN: Yeah, you've got a long left tail. You've got lots of guys down here and that's okay. We know about some teachers with more precision.
MS. FEILD: I think that from a policy perspective, right, we have to think about it from the committee's perspective. If we choose

a model that looks at two lags, three years of data, then we have to be able to tell teachers fourth grade teachers will never have that precision. We would never be able to tell them?
DR. DORAN: No, no, there are other factors. Two lags brings additional precision. There are other things that yield more precision as well. So the standard errors of the teacher effects depend on a number of different things. So, for example, it depends on the number of kids within a teacher's class. It depends on the homogeneity of the students within that class. It depends on the number of lags that's used. It could depend on some of the fixed effects. So there are other factors, yes, those grade four teachers will never have a prior lag as the grade five teachers and up would. But that's not the only thing that determines what the standard error of the teacher fixed effect is.
Jon, you wanted to say something about the fixed effects, right?
DR. COHEN: Well, we very quickly started focusing on the covariate model. The simple differences model, Model 4, displayed on the

top, you see the standard error is there. I'm eyeballing this. The standard error is around 40 or 50 points. If you'll recall from this model, the estimated effects were also larger. The typical effects were estimated to have larger magnitude, either negative or positive, but larger magnitude. So fifty points relative to the size of the effect being estimated is not as bad. I want to make sure that we don't ignore this model unless the committee wants us to ignore the model.

I think the -- you can see why -- let me back up a bunch of slides and show you this model is finding bigger typical effects. It's because we're estimating something different. It's estimating deviations from a different line. Let me go back there and show you that again.

Remember the red line here, this is a simple differences model, and this is the one that shows you what is typical, what is currently typical. The magnitude of the effect, a typical student -- a teacher teaching only typical students with typical growth right here will appear to have a big effect. I don't know

how big that is on the scale. That's probably 200 points on the scale down here just for teaching typical students, so there's that big gap. Up here you also have that similar sort of advantage, just not as a disadvantage in this particular range. You see more variability. See estimates of effect that have a greater magnitude because of deviations from a different set of lines.

Does that make sense? Okay. So now we can leave this model for as long as you guys want to leave it.

Next slide.

DR. DORAN: So moving on to the next criteria, we don't want to evaluate any model in isolation. Personally, you might care about precision more than you care about personal. Those are your own opinions. Sam's going to facilitate a conversation later on where you get to express which of these criteria you place more weight to. This is up to you, all right, but don't evaluate (inaudible) -- you get a model on one of these criteria. So why don't we move on to the next one unless someone wants me to stop here and explain more or have larger

discussion on precision? Yes, Stacey?

MS. FRAKES: Model 4, the distances model, is that correct?

DR. DORAN: Yes.

MS. FRAKES: And the others are the covariates?

DR. DORAN: That's correct. So Model 4 is the simple differences model and all of these are the covariate adjustment models, and Models 1 and 1A are the ones that have the teacher effects and anything with a 3 -- here's an easy way to remember it -- anything with a 3 is a 3-level model. It has students, teachers, and schools.

MS. FRAKES: Thank you.

DR. DORAN: Jon, is that why they were called Model 3?

DR. COHEN: I'll exclaim yes but I might not be telling the truth.

DR. DORAN: I actually just thought of that now and I thought, wow. Just because the way we write them and the analysis specifications.

DR. HOVANETZ: I will have to say, though, we did revise them to have simple differences, that's the reason why a two tiered model is 2A

and a three tiered model is 3A.

DR. DORAN: Well, it works out to be a beautiful heuristic device. If you have a 3 next to the model, it has teacher, school, and student effects. So that's why I prepared that little summary sheet for you so you could follow along because we're going to refer to these models by number throughout the course of today and tomorrow.

Yes?

MR. LeTELLIER: Real quickly, the regression line that you just --

DR. DORAN: You want to revisit it?

MR. LeTELLIER: I don't know if we need to revisit it. It's on page 11 you just showed, but it shows the slope is straight, and my question is if you were going -- if a kid one year was 1,400 and the next year you're expecting him to be about 1,450 or something like that, if the kids in that area had a specific number that you're expecting them then would that line be straight or if you magnified it would it kind of wobble as it's going up?

DR. DORAN: Well, the simple differences model forces the slope of that line to be fixed at one, okay. The covariate adjustment model takes all of those idiosyncrasies across and finds the line, the slope of that line, that best fits the data given some of those wobbles because you do expect some wobbles. Now you can do other things statistically to account for some of the curves and the data and you can fit a whole bunch of different things, but that line finds the best fit of that line through that cloud of data.

MR. LeTELLIER: Okay. I guess I was having a hard time with that because it just seemed like we're making that slope and now we're just -- all kids are following this where you may have one area of growth that may be higher or lower, and so we can't really account for that precision.

DR. COHEN: Well, we do in fact, and the other model where we could -- let's go back to that. We don't have a graph on this, but with the help of a graph we do have a bit of imagination as an interpretive dance. All right. Ignore the red line, ignore the man behind the counter, ignore the red line. We'll go to the green line, it's just a straight line.

Now let's suppose we include this covariate a little flag which we did, a flag that says this is an ELL student, a student who is in one of his first two years of ELL. There will be a separate line for those kids. It will be parallel to this line and in fact is parallel to this line and a little bit above this line. Then we had another dummy variable for students with autism spectrum disorder. I'm trying to remember; I believe that one actually had a positive intercept, also. That will be another parallel line that is a little bit above this line. I don't try to remember all the details. A kid who's been absent for 60 days will have a separate line that's again parallel but dropped this much. Katie's been absent for, I don't know, more days than that, but as per 100 days might have a line down here. So the covariates actually account for that sort of -- those sorts of differences among students given their measured characteristics. It doesn't try to fit just any individual differences in the data.

MR. LeTELLIER: Right, I understand that. I remember that from last time, if you're looking at the X and Y access the kids that are

at a 500 versus 2500, and if you go in that range all kids are going to fall along that line and that may not be precise at some point. Like, maybe the kids that are at 750, you actually would go up a little bit.

DR. COHEN: You can see the scatter plot here. That's why we have the scatter in the background; you see it's pretty evenly distributed there, right? You see a little bit more variability down at the lower end than you do at the top end and a little bit less variability, but it cuts through the center of that cloud. So often what you'll see if the model doesn't fit if you have that kind of problem is this cloud won't really be centered. It will look more like a curve or a curve the other way and we don't see that.

MR. LeTELLIER: If we expected kids to be -- a kid's that at a thousand this year, we expect that he will be at a thousand next year in his next grade or wherever this line falls?

DR. COHEN: Whatever, yeah.

MR. LeTELLIER: Right. So I guess what I'm saying is that -- 'cause last time when we had discussed this, we were going to look at where

students were in that range, so similar students at a thousand where they would be expected to be. Couldn't there be a -- see, I guess maybe mathematically I am not going with your statistically. Couldn't there be that thousand range, all kids that were around a thousand maybe generally they were at 1,100. The kids that were at 1,200 maybe all those kids scored 200 points higher and the kids that were underneath, let's say at 750, maybe they went down to a 720. That line can't take that into consideration because that's a single straight progression line.

DR. COHEN: That is right. We did look at some models that had curves in them and they didn't fit the data. We wound up seeing -- you'll notice there's a mass of kids up here who are getting the highest possible scores and you -- you see highest possible scores; you don't see as many lowest possible scores. It's not as clear in here. But when you put the curves in then those start to drive the analysis, and so we wound up -- we actually did see residuals that didn't have that nice even pattern. So you had prediction errors which would translate into

mistaken attributes of teacher effect.
You saw prediction errors that weren't constant across the range. So we tried and we didn't see that, I guess is the answer. I mean, there are literally an infinite number of curves we might try fitting. You know, we tried some.
MR. LeTELLIER: Why does it have to be -- I don't know if I'm stating it -- why does it have to be a line? Why can't it just be whatever the students in that area --
DR. COHEN: Well, that's exactly what this is. This says within the -- and the areas are defined by achievement level. It says, okay, kids in achievement level one we're expecting -- I don't know what the number is -- a hundred point spread. Kids in achievement level two we're expecting 50 points. That's the simple difference; that's what the red line is doing.
MS. BOURN: Can I take a stab at what I think you're asking? I think he's asking why can't you look at each score point along the axis and take the average of what the Y-axis would be for the score point and that's the expectation, instead of just using a line to describe the entire set of data.

MR. LeTELLIER: Yes, that was much more eloquent.
DR. COHEN: I suppose we could certainly do that.
MR. LeTELLIER: It doesn't make sense because what we were discussing last time, and I'm just throwing this out here, because what we were looking at, and I remember clearly conversations because I remember the number 140 offhand, and we were talking about students next year, we were going from a baseline, those students would be -- the teacher would be assessed according to where the general kids at 140 were the next year. And with a straight line with that big of a spread, there's no way to do that, and in the center of the line obviously as you moved the slope there's less movement and the outer parts move more.
DR. DORAN: Can I ask for a clarification? Are you asking what would happen if we took a particular range of the X axis here of those kids that were in that range what the expected score on the Y axis? Is that what you're asking?
MR. LeTELLIER: Yes.

MS. BROWN: He's also asking for multiple regression lines or regression line per --
DR. COHEN: It wouldn't be a regression line. They're asking for each scale score point, right?
MR. LeTELLIER: So it would look like this? You have a point at 1,000 and 1,100 and --
DR. COHEN: You could do that when we estimate a model like that. What we do is we just have a separate point for each scale score point. What you give up there is you give up precision, the precision of the expected value and therefore the precision of the teacher effects would be much, much, much more variable. You've got 70 or 80 different scale score points, so you'd be dividing your sample into many smaller units. So you would give up precision, so you might see that this scale scores where the average is about there and this scale score point is here, but they might not even be any different. We can try something like that. That hadn't come up in the prior meetings as something to try.
MS. GINN: Why would that be important if it was used?

MR. LeTELLIER: Well, because and I think it's a matter of looking at statistics versus mathematics, and if I'm the only one here that's fine; I don't want to beat a dead horse. But I'm thinking that more people are on this line than not just -- in real life a kid that scores 1,000 that next year scores 1,100, and we say most of the children in that category that were 1,000 last year are now 1,100; that's the expected growth. That's real life versus this line which is not going to take that into account, I don't think, as fully from what --
PANEL MEMBERS: (Over-speaking.)
MR. COHEN: Let me show you another fairly simply graph that we have that we just created.
DR. DORAN: While he's creating that, I think that there may be just a little misunderstanding. Either it's me or we're just talking past each other. What you just said now was in the real world there are kids who have 1,000; and there should be some prediction that's similar for those kids who had a score of 1,000 the prior year, right?
MR. LeTELLIER: No, what I'm saying is real life scenario, this year they're at 1,000; next

year we'll see they're at 1,100. We're going to see those two numbers. So when we're looking at outcome and we're saying where a teacher should have their students, we can visibly see that the general student that was at that level should be at such and such level.

MR. DORAN: That's what this is doing, that's what this is doing. This is showing for any given -- for any kid who had the same prior test score, what are they expected to do in the subsequent year? That's exactly what this line does.

DR. COHEN: There is a linear assumption here that that data goes in a line. He's saying couldn't we do something non-parametric where the line and plots go wherever it wants to.

MR. LeTELLIER: Yeah, I understand this plot. I see where the kids were the first year or the second year, but life is not linear.

DR. COHEN: It's often more linear than you think, but let me -- I have a question over here.

MS. BROWN: Well, I just want to see if this is where you're trying to go because it isn't linear when you look at each case in

point, and that's absolutely true. The purpose of this is to look at a large set so that we see on average because those -- if I'm understanding this correctly, all the points are actual performance, and that line is not a forced line. It's not a -- we have determined by some calculation that the line should be here because that's what we should expect of students; the line is actually where the average falls. So in a sense even though it is linear because when you draw a progression line through a scattered plot of points it will be a linear line, but this is actual performance and it's attempting to show the average over a large set of data.

The problem is going to go back to kind of where you were talking before. What if it's only that one kid, and that one kid really makes a difference for that one teacher?

MR. LeTELLIER: Well, yeah, let me just draw this.

MS. BROWN: Go, John, go.

MR. LeTELLIER: I'm just going to use three numbers here. What I'm saying is this: If we draw a line that goes like this, and I totally get that; what I'm saying is what if the kids

that were right here on average actually score here; that would make it here. The kids here on average score here and the kids here on average score here. When you connect those points these are still average. Again, that's not what I'm saying.

DR. COHEN: Let me show you what you would see. If this were the case you would see that on that graph.

MS. BROWN: You would see a plot line.

DR. COHEN: Let me draw you what you would see. If we try to defend it around this line, you would see a scatter plot that looks something like this.

MS. BROWN: So the scatter plot would have to be shaped --

DR. DORAN: You see that plot in the residual and you don't see that, you don't see that here.

DR. COHEN: Let me see if I can -- I have some data --

PANEL MEMBERS: (Over-speaking.)

MS. BROWN: Can I just say that in like common terms?

DR. COHEN: Yes.

MS. BROWN: Because now you're both saying the same thing and the point is if, in fact, the actual performance of the average group of students was the line like this, the scatter plot of points behind the line would also be a shade like this.

DR. COHEN: But the line will still be straight.

MS. BROWN: So in reality if actual performance is what you say, if there's a year and there's an anomaly in the scale of the test or the scoring or whatever, then the regression line would be funky like that because that would be the actual performance behind it.

DR. COHEN: Yeah, and it would be detectable. You look at the scatter plot and you say, you know, Jon, that dip there maybe you should be pitting a line with one person.

MS. BROWN: But typically this is what you see in this data set and it is feasible that that could happen.

DR. DORAN: There are data that we're not presenting that I can take over time on the computer and show you; if supposing the world was exactly as you were suggesting, Jon. We would see the scatter plot like that, okay? Now we compute what's called the residual. The residual was the difference between the kids' observed score and their expected score.

We did many, many analyses to look on those residuals to look at whether they fit particular patterns. They normally distribute it, how did those fitted values plot -- the fitted values plot against some of the residuals? Those strange trends would have shown up in our analysis of the residuals, and we would have thought we need to go back and fit cubic trends, quadratic trends, do some local smoothing, you know, use different polynomial terms to account for the bends in the data. We actually did do something like that back at the lab to see if those models fit the data better than these straight line models and they didn't. Not only did they not, but in all of our analysis of the residuals that we do that we're not presenting here, but if you want to see we can show you some of these things, we don't see those strange patterns of the curvature that we would get to.

Perfectly reasonable what you're asking to look at whether that's true or not because if we

do see that the role is not as linear as these models have turned out to be, because it's just a line, a straight line -- okay, perhaps the world isn't as straight as we made it here. We can test those assumptions, we did do some analysis. I can show you some of this stuff if you want to come over here to the computer and look and see.

DR. COHEN: Just for public record, the world is round.

MR. TOMEI: We're never going to be able to develop the perfect model. You develop the best model you possibly can and then to handle the extreme outliers you have to have some policies or business practices in place, you don't evaluate a teacher, you don't have a minimum end of student population and you don't make high stakes decisions about teachers based on a single year's worth of data. Those types of things will mitigate the kind of concerns you're expressing in ways that the best model in the world would never be able to capture 100% of the time.

DR. DORAN: There's a very famous statistician; his name is George Fox and he said

all models are wrong, some models are useful. We can't fit a model that fits, you know, into every single one of these points. That would be what's called a saturated model and it's not a reduction of the real world. So we have to investigate which model is the best reduction of the real world data that we can apply and get reasonable inferences from. So that's what we're trying to do here. That's why we're looking at other fixed effects, other covariates and control variates. Adding those terms when they fit the data and then when they're excluded.

All right. Christy gave me the flag of five more minutes. Why don't I start on something and we'll finish it after --

We're going to start with spool effects now. We're going to define why we care and what we're looking for. So again, using the same questions that we looked at before, let's look at the question.

Should value-added models account for systematic differences between schools? The question is a straightforward question but it's a hard one to answer. Remember, some of the

value-added models include school effects, some of them include only teacher effects. Let me talk about why that's important and what matters, all right?

When we think about school effects and teacher effects, a teacher effect is essentially in the world of value-added modeling the things that we think that teacher has done to cause the improved or not improved student levels of achievement. However, suppose there are other things happening in the school that are also cause for students to improve in their learning or not improve in their learning that the teacher is not responsible for.

Suppose there are qualities of the programs initiated by the principal or the curriculum coach or some other advocacy group within the school. There are things that are systemically happening in the school that are a cause or at least part of the cause of student change that are outside the control of the teacher. If you do not include school effects, those changes, those things have to go somewhere; and if you ignore them, they get pushed into the teacher effect. What would end up happening is if you ignore school effects and let's say there are good things happening in the school and let's suppose those school effects are real and big, what happens is they get pushed into the teacher effects and those teacher effects may appear to have higher added value rankings than they truly deserve because they're getting some credit that the school is doing.

I look really good because things that the school does that I didn't initiate got pushed into my effect or vice versa. Things that the school is doing are dragging down a particular teacher's effect. All right? So that's part of the reason that we're interested in looking at this. Are there things at the school level, initiatives or programs that are also partly because of changes in student growth that we want to account for? When you do that then what you do is you partition the variability and growth. We get teacher effects and we get some of the variability and growth that's due to the school. Those are estimated.

So what do we look at? That's the question. Should school effects be included? What statistic do we look at? We're going to

look at the variation in student growth between schools.

Evidence in favor of a desirable model. If the model suggests that there are systematic school effects then the policy must decide how much to attribute. We don't need to answer that question today; the question will be: Should school effects be included or not? I don't think that's the goal; we're not going to answer that, but you can attribute none, all, part, some fraction of that. The issue here is should we include school effects in the model or not?

Why? As I said, determining if and how much of the school effects should be attributed to the school, then also if the school is doing something that the teacher is not responsible for and you exclude school effects, teachers may appear to be higher performing, higher value added or lower value added than they may truly deserve.

Let's look at something and think about this as we get ready for lunch. This is the only graphic I'll show you before we get ready to -- I thought we'd finish our slides by 1:00 today. I was way off.

These are the teacher effects on a standard deviation scale, all right? Essentially, what this means is this is Model 1, looking at your sheet you'll see that Model 1 is one that includes only teacher effects. What we see here is that the teacher effects appear to be really big, okay. The number 30 is not necessarily meaningful saying ten points higher than this model over here is irrelevant. What's important here is that you see the magnitude of what's happening here. So when you have teacher effects, teacher effects appear to be really big under this model and this model.

Now here in Model 3 we include teacher effects and school effects. Note what happens here. Teacher effects get smaller almost always, always relative to this, teacher effects get smaller and the school variability -- the school effects -- are relatively big. What is this graphic suggesting? This graphic is suggesting that if you ignore teacher effects in reading, if you ignore school effects in reading, teacher effects appear to be very big. When you then account for school effects and teacher effects, that is, those systematic

things that schools also do that may be outside the control of a particular classroom, but also are the cause for a relationship with an improvement in student achievement or changes in student achievement. Schools have an effect. They seem to matter.

So essentially what you're doing is now you're saying, well, we've got those teacher effects but those teacher effects really aren't as big as we thought. Some of what's happening is because of the school. So we pull that out of the teacher effects. That's why the teacher effects are not as big under these models as they are here.

Let's just look at the next subject and then we'll leave and explore this issue when we get back. The next graphic is the same but from that, and we see a similar pattern. We see that when teacher effects are included only the teacher effects appear to be very big, but when school effects are included alongside teacher effects, they appear to account for a large proportion of the total variance and scores and teacher effects get smaller. Let me explore one thing with you.

Let's pretend for just a moment that we included school effects but suppose this part of the school effect was really small and this was really big, okay? And this was small and this was small and that was small. That would suggest that even after accounting for school effects, teacher effects are still really big. If that were the case, we might say school effects are not necessarily needed. Now I can't tell you how big the effect of the school should be before you make a decision, a policy decision on whether or not you include it. This is data, these are data that suggest that school effects seem to account for a large proportion of the differences in student achievement. Maybe schools matter.

When we come back from lunch, Kathy, I'll let you close this out soon, we'll explore the observations that we're making about the differences between models and whether or not you think school effects should be included or only teacher effects, and we'll talk about any questions you might have.

MS. HEBDA: Thanks, Harold. A lot to chew on during lunch. We're going to stop the

recording and break for lunch now.

(Whereupon, a lunch recess was had.)

DR. DORAN: Before we went to lunch, we started with the question of whether school effects should be included in the value-added model, in addition to teacher effects, or only teacher effects. Let's go back a step and talk about how do you do school effects, or what is a school effect? Is it a certain set of variables that you put in? Let me try and describe this in a way that is relatively straightforward to understand what the school effect is. But let me start with just the teacher effect and then we can generalize what the teacher effect is to the school effect.

When we estimate the regression model for teacher effects, essentially what we get is we get an expectation for all kids. We talked about this last time. We said what's the expectation for kids if they have a typical teacher or a teacher of average effectiveness? When you estimate a teacher effect, what we look at is Sam had some group of students. We look at the students in his class and how they deviated from that typical expectation. They

all have, let's assume, very high positive deviations from the typical expectation. What's the expectation? Remember that regression line? That's the expectation. Every kid in your class, Sam, had positive deviations from that typical expectation. So now we've got what we call a residual. It's a little bit more than just taking the average, but let's just work with this.

So now I've got for all the kids in your class, I take the average of those residuals. They're all positive, which means every single kid in your class for the sake of argument meets the typical expectation. Let's just say they beat that expectation by a lot. Your teacher effect would essentially be about the average of those residuals. They're all positive, you look really good relative to the typical teacher in the state.

Now I've got another teacher and so suppose now I have a group of students and I didn't do as well with those kids as would be expected. Suppose each of the students in my class fell below that typical expectation. Now I've got all those negative residuals and we take the

average of those in a sense, and now I look very low relative to the typical teacher. What's interesting is the teacher effect, you have those residuals the difference between how they perform and how they were expected to perform, and you average them in a class to get that for teacher one, teacher two, you get that for every teacher in the state.

Essentially, what the teacher effect is, the teacher effect is the deviation from -- the kids in that class -- is the deviation of the students in that class from the typical expectation. So we know what the expectation is, kids in a teacher's class in many respects -- kids in that teacher's class had in effect a deviation of plus ten. It just means maybe they scored ten points higher on average than the typical student, than was expected, or a negative teacher effect is one where the teacher has -- let's just call it negative ten. That means students in that class performed lower than was typically expected. That's the idea of the teacher effect. It's a deviation from the expectation for the kids in that class. You have positive deviations for all your kids; I have negative deviation for all of my kids.

Now the world is not that straightforward. Your class maybe some of them have positive deviations, some of them did not, so there's a little bit more than that. Again, you still take all of those residuals and we do a little math with those and we come up with, well, for the most part your kids did pretty well and for the most part my kids did not do as well. So the teacher effect is the deviation from that expectation.

The same concept applies to the school effect. So essentially what we've got is a group of kids who are within a school, from many schools, and essentially what we're looking at is those kids, how did they fare relative to the expectation? All of the kids in the school? Then kind of in this -- now when we actually do the math, mathematically now what we do -- now I've got this effect. See, all of the kids in this school on average perform lower than was typically expected. Now I look at, well, the kids in Lance's class, they're in that school; how did they perform relative to the school average? So now I've got kids in your class,

they performed higher on average relative to other kids in that school. Me, kids in my class, did poorly relative to other kids in the same school. So now I've got two deviations for a school effect.

I've got an expectation, I expect all kids to learn. Then I've got a school effect. All the kids in the school, how did they fare on average relative to that expectation? The school effect is still a deviation just like the teacher effect is, but now I've got a second deviation, now I've got teacher effects. How do teachers now deviate from that average school effect?

How do you do this? It's a statistical thing, all right. We essentially see that it's kind of like the way I described it. Essentially what we get, and again, it's a little bit more than just an average, but essentially what we get are the residuals for all kids in a school. That's the school effect. And for the teacher effect, again, it's the residuals for the kids in a class and we see how far they deviate from that school effect. That's what a definition of a school effect is.

It's the deviation. Then the teacher effect is the deviation from that school effect. Yes?

MS. ACOSTA: I have a question about the average that we get from -- including everybody. That includes students within that teacher's class so those kids are counted within the average and then separately? Do you see what I'm saying? As well as for all the things we're looking at. When we say we're going to get their regression line for everyone, that includes for example students with disabilities. So we make another line with just the students with disabilities; so they're in the initial line as well?

DR. DORAN: That's right. So let's just play it out a little bit. We've got expectations for kids with disabilities and some of those kids are in your class, so we've got observed scores, and we say did they beat those predictions? And they did and there are other kids who don't have disabilities who got a separate expectation for them when that variable is included in the model and did they beat that expectation? They did. So no matter who you taught, all the kids in the class beat their

expectations. Now we've got those residuals and we'll average those together, say, and they did better than expected.

So where we were going with this is to these slides and asking the question should the value-added model include both school and teacher effects? Part of the reason that this is interesting to look at is because when you look at this here we see that teachers appear to be very different from each other. There are some teachers who would appear to have high value-added and some teachers who appear to have very low value-added. There's a big spread between the teachers. Okay?

But when we include a school effect what happens is maybe those teachers appear to look good as a function of something else that's happening systematically in the school. So when you include the school effect the teacher doesn't appear to be quite as good -- almost as good but not quite as good because some of the things that were being pushed out into the teacher effect are now accounted for by the school. So it's going to change and that's what we see in these graphics here. We see that when schools are not included teachers appear to be very different, but when schools are included they seem to account for some total proportion of the variance. It's not tiny, it's some portion. And teacher effects are smaller than when you only have teacher effects.

John, you had a question?

MR. LeTELLIER: Yeah, the counter would be true as well if you had a negative school effect, then that would change the teacher effect as well?

DR. DORAN: It could.

MR. LeTELLIER: Okay. I'm not saying that's going to happen, but that could happen?

DR. DORAN: It could happen.

DR. COHEN: Yes, it does happen. Some schools are better than others -- not better. Some schools show a higher student effect, some schools show lower student effects. Can I say something else?

DR. DORAN: Yeah, go ahead.

DR. COHEN: This is pure policy decision. We're presenting the statistical data and the data tell you go from policy decision to make.

This line here shows you that student

growth varies somewhat by school, not as much as it does by teacher, but it varies by school. Meaning that if you take it out of the model, any gross that's common within the school will be attributed 100% to the teachers, and that's fine if that's the policy decision you want to make. Model 1A and Model 1 make that -- it's a policy decision. We've decided that the only thing that matters at the school in terms of student growth is the teacher effects, and so any school level effect that may exist is really just the net effect, the aggregate effect of having all good teachers.

So if you want this model, the policy decision you want to make would basically be saying that yes, it's true that more effective teachers are clustered in some schools. So what? They're still great teachers. These less effective teachers are clustered in other schools and, so what? They're just less effective teachers. These models -- they're telling you that there is correlation in teacher effects within schools. Some schools have higher student growth, some students have lower average student growth. Your policy choice is

whether to attribute it all to teachers, whether to attribute none of it to teachers, or whether to attribute part of it to teachers.

MS. GINN: Question. Then really does it show that without putting the school in teacher impact has allowed fluff?

DR. COHEN: That might be a technical term there, right?

MS. GINN: You know what I'm saying.

DR. COHEN: I can't make the decision for you. This is not a statistical decision.

MS. GINN: I'm just asking you, based on what you have there because of the school impact, we'll keep the difference, okay.

DR. COHEN: All I'm --

MS. GINN: We don't want to touch that, fine. Here's my second question.

We keep using the school effect positively, but the school effect can also be negative.

DR. COHEN: By definition the models are estimated, half the schools are better than average, half the schools are worse than average. So it's positive or negative. It really has to do with what you believe affects -- if you believe that any common influence in a

school is 100% student teachers and you're ready to stand by that, choose one of these models or use one of these models that has the school effect back into the teacher effect.

MS. HALL: How are the schools being rated? Is it the same thing as the teacher or is school grades or how are we determining -- how are you coming up with that graph?

DR. COHEN: Oh, the school effects are estimated the same way. I am going to speak loosely, this is not mathematically how it works exactly, but it is this general idea. It gives us a -- if I were to take these estimates that are only teacher effects and find the average effects within school, I would then subtract it off the teacher effect, put it down here and make it the school effect. So it's kind of -- if you leave school out of the model, it's kind of the average teacher effect.

MR. MURPHY: If you don't remove the school effect, you have teachers that may or may not be earning that additional portion.

DR. COHEN: Yes, it's important that you recognize that as a policy and substantive decision. If nothing statistical models can tell you, it's not -- sure, Harold or I would be happy to tell you what the right answer is, but not statisticians and not giving you an answer. This really has to do with what you believe matters in an education system and how you think accountability ought to be distributed between leadership and the teachers.

DR. DORAN: This decision would be really easy. Jon's exactly right; this is a policy decision. Remember, we looked at precision before. Everything is policy decision, slash, statistical decision. But here if we did this and we said that schools didn't account for anything at all, your decision would be relatively easy. There's no need to account for the school effects because they don't account for any other variance and scores. But here we don't see that to be the case. We see that schools seem to have differences between them and that that seems to affect the differences in the teacher effects, right? So there is something that matters systematically between schools.

Yes?

MS. FEILD: Let me make sure I understand.

The school effects here is really the same as the teacher, but it's kind of aggregated for the school?

DR. DORAN: All kids, yes.

MS. FEILD: So what happens is, and I think someone may have asked this question, you may have a school who was rated under our accountability system as an 'A' school, right, based on points? And it could have been proficiency, gains, whatever it happens to be. Now you have this model where the teachers performed okay but not great, but based on the nature of the statistical model where half are going to be below the average and the other half, this school who is an 'A' may have every teacher get a negative school effect on their individual evaluation; is that correct?

DR. DORAN: No, it depends. We're going to get to the end of the day -- now, if we use the average as a classification rule, you say -- classification is something that we're going to talk about at the end of the day today. Suppose we say that any teacher above the average if it's called a high value-added, any teacher with a below-the-average is a low value-added.

That's a classification rule and now how those effects relate to the other grades, I have no idea and I don't want to speculate on it. Whether it's plausible that they could look good under one model and not look good under another maybe, but there's no basis for me to make that decision. That's pure speculation.

But the issue about classification is something we'll talk about later today; and that is do we have to use the average? No, you can use above the average to create more stringency and go up by five, one, you know, you can be one standard error above that, you can be two standard errors, one-and-a-half standard errors. We can come up with different classification rules and we're going to show you the consequences of what happens when you choose the classification rules at the end of the day.

MS. FEILD: I think you misinterpret --

DR. DORAN: But how we score grades, there's no way to --

MS. FEILD: I understand. I think the misinterpretation is when we talk about a school effect, at least in the original conversation before lunch, I was thinking we were looking at

other variables that had to do with schools separate from value-added, school grades, accountability, but in --

DR. DORAN: No.

MS. FEILD: But in essence we're talking about the same thing. Okay. I just want to make that clear.

DR. DORAN: That's a good clarification, all right. So essentially, another way to think about this it that we've got -- this is all value-added, all the same metric, and this is everything that we're looking at. So essentially we're going to have some expectation and now I've got -- I'm just drawing -- it's too low. I'm a school who on average beat the expectation and now I'm going to have teachers in that school, some below the school average and some above that average by some amount. It's all based on the same data, even the same model. We don't have other variables like school grades and so forth.

Now what's interesting here is that we see a couple of things, schools appear to account for some non-trivial proportion of the total, variation in scores. Non-trivial is not determined statistically. We've just shown you the graphs and just based on the visual displays it seems to be a large proportion of the total variance of scores. Teachers are less different from each other when including school effects. What does that mean?

When we don't include the school effect, we have this much variability between teachers. There's a lot of variability. Teachers are very different from each other, but when you include the school effects part of what made that teacher look really, really good before gets soaked up by the school effect; or part of what made that teacher look really, really, really low value-added before gets soaked up by the school effect. The teachers appear to be less different from each other under that model. Yes?

MS. FRAKES: So when the school effect is smaller, the teacher effect could be greater. So when you're talking about --

DR. DORAN: It would be greater by definition, so as this gets smaller this will get bigger.

MS. FRAKES: So when you're talking about

moving your best teachers into your lowest performing schools, I mean, when you're taking into account as a school this is the school effect that the students are not performing, and then that teacher goes there and she does what she does so well; that's really going to show as opposed to that school effect, correct?

DR. DORAN: That's correct. The teacher would appear to be very high performing within that school.

MS. FRAKES: Within that school because the school effect is lower. I mean, it can work both ways.

DR. DORAN: Correct, if you've got a bad teacher going into a good school that teacher would have a lower value-added effect within that school than they would in some other situation potentially.

MR. LeTELLIER: So could this have an unintended consequence of a performing arts school, which typically we have data that shows that those are higher performing schools, or some sort of a magnet school where, you know, an IB school or something where the kids are very well performing. Will that make it harder for a

good teacher to have a good value-added model score because those kids are already performing so high, the school is performing so high, the teacher goes in and the kids are doing well.

DR. DORAN: In some respects, yes. In some respects that is almost what you would expect. If those kids are doing well because of things the school is doing then you want to differentiate that from the teacher effects, right? But if you ignore the school effects in that particular case then everything that's happening systematically in that school gets pushed into the teacher effect, and those teachers may appear to be high value-added, not necessarily because of what they're doing but because of other initiatives in that school. But now you account for the other things happening in that school, and the teacher effects could -- in fact, they will be smaller as we see here. The teacher now has to demonstrate that they're doing things above and beyond what's normally happening in that school to have a higher value-added effect relative to other teachers in that same school.

MS. BROWN: I think what is discomfort and

where some people are settled is how then when you look at the full distribution is there still equal opportunity to achieve a, quote, if 50 is high -- just pick a number, doesn't matter -- can you -- do you have equal opportunity to achieve a value-added score of 50 at a school with a more positive school effect as you would at a school with a more negative school effect because if you are the same teacher in both locations, implicitly you should have the equal opportunity to achieve the same score.

DR. DORAN: Like if you're a good teacher -- let me just pull away from the number 50 for just a moment. If you're a good teacher and you're doing things above and beyond what other teachers in your school are doing, whether you're in a good school or bad school, you have equal opportunity to appear as a high value-added teacher in either instance. Remember, what you have to do here is those teachers have to do more than what other teachers are typically doing in that same school.

MS. BROWN: I understand.

DR. DORAN: So that gives them -- so long as they're doing those things, they have equal opportunity.

DR. COHEN: Can I break in here?

DR. DORAN: Yes.

DR. COHEN: It's a little more complicated than an extra decision that you have to make. I'll put a little example up here. Let's say we're going to talk about one school. That school has four teachers, and the math is not exactly -- this math does not come out exactly the same, but it comes out really close in this illustration.

So the school effect, estimated according to let's say Model 3A -- school effect is 10, teacher effect is 20 minus 20, 10 and minus 10; so they're equal. They have an average of zero around the school mean. If we were to ignore the school effect and just estimate it for the teachers and say everything is attributable to the teacher, those scores under Model 1A would be 30 minus 10, 20, and 0, right. So this is attributing all effects to the teacher, so we just assume that any school effects are just the result of an aggregate teacher effect.

If we take 20 minus 20, 10 and minus 10,

that's attributing 100% of the school effect, of whatever is common in the school to the school. So great principal or a bad principal, it's all on the principals. But what we did down here is we say, well, we decide we're going to split this half and half. We're going to say half of it is due to the fact of the school and half of it is due to the teachers themselves. Well, then we wind up 20 minus 15, 15 and minus 5.

The thing is if we estimate it with the school effect, this decision is explicit. Do I add these things together and get that? Do I partition it partially and do that or do I attribute it 100% to the school?

MS. BROWN: However, if you look at that differently and you use your same numbers, I think the question around the table is if school A has a school effect of 10 and school B has a school effect of 5 yet the same teacher effect of 20 in each of those schools -- or maybe I'm backwards, but the point is that is if we went that way, is it possible that in the long run teachers at the school effect 10 have the potential to have the higher scores? And if there was a lower school effect, would those

teachers potentially within the district or within wherever have lower scores when compared to each other?

DR. COHEN: You're unwinding it. You're making an assumption about what you're going to do with the school effect. If you ignore the school effect, you wind up with this. If you have Model 1 or 1A, you wind up attributing anything that might be the school effect to this.

MS. BROWN: I understand that.

DR. COHEN: Right, and that's the situation in which if there really is an independent effect of the school that is not due to the teachers, then the schools happening to be in the better schools will be advantaged and the teachers who happen to be in the worst schools will be disadvantaged.

MS. BROWN: Correct.

DR. COHEN: One might argue, and I'm not making this argument necessarily, but just to point out, somebody might say, yeah, the principal matters but he does all of his work -- he matters through the teachers who are interacting with the kids day to day. So the

principal's influence is through teachers. Therefore, it really is a teacher effect. Someone could argue that this is average teacher effect and that's how you should do it.

The point being that there is a policy decision to be made. The data won't give you any information to help you make this policy decision; it's how much of what's common within school gets attributed to teachers and how much gets attributed elsewhere.

DR. DORAN: Let me try one other angle. Suppose you have two schools -- let me try this another way. We're going to try this.

MR. MURPHY: Teachers are being compared in the same school, not across the district, so it's all relative at the school level, not at the district level.

DR. DORAN: That's correct.

DR. COHEN: But only if you attribute it -- you can add the school or a portion of the school effect back in once you have the estimate added in, then you can attribute it however you like. So it doesn't have to be relevant. That's the policy decision.

DR. DORAN: Here's another thing. I think this is part of what you're getting at. Suppose I had two schools or two teachers and suppose some teacher had a -- suppose at the school A it's typical for the students to score 20 points above the state average. Let's just use that term, it's typical for kids in school A to score 20 points above the average. Now suppose that that teacher, a teacher within that school has kids who only score 20 points above the average. That teacher didn't do anything above and beyond what's typical for students in that school, right?

Now let's take another situation, another teacher, who's in a different school. In that school, let's just say it's typical for teachers to score 10 points below the average. That's typical for that school. But that teacher, a teacher in that school scores 20 points above the average, the same as the other teacher, okay, in the different school. That teacher would appear to have a higher value added because that teacher is doing something that's very different than what's typical in that school. Do you see? So while two teachers did the same thing, this teacher over here is doing

what's typically expected or typically observed for kids in that school. There's nothing exceptional happening there, whereas this other teacher went away above and beyond what's typically observed for kids in that class.

Now while two teachers have the same -- both have kids that performed at the same level given that they're in different schools this teacher did what's normally observed -- or did what's normally observed in the school, so it's not regarded as being high value-added because that's what's typical for that school. In that instance you could have teachers who do the same thing with their kids, but this teacher over here is doing something that -- that's what you observe with any other kid in school. So what makes that teacher particularly high value-added when the school effect is included? When the school effect is not included, both of those teachers would appear to be exactly the same. They both beat the average by 20, right? But unless you know something more about the context of the school that they're in, well, this goes to -- that's what you typically observe because there's other things that the school does to get

the kids there.

MR. MOREHOUSE: What happens to the students who perform below the average?

DR. DORAN: Again, a teacher has a group of kids who perform below the average of the school?

MR. MOREHOUSE: Yes.

DR. DORAN: They would not appear to be high value-added relative to that school. They don't do -- that's what we regard as a low value-added effect because --

MS. BROWN: But don't confuse that with a teacher that has a group of level 1 students. Those are two different things.

DR. DORAN: That's exactly right.

MS. BROWN: Because you can have all students that score at level 4 and they all don't gain, and you would be considered (inaudible). But just because you have maybe lower performing students coming in, that wouldn't necessarily be the case.

DR. DORAN: Right, right, because of intake. The kids you intake are not to be confused with the potential for a value-added effect to be higher.

MS. KEARSCHNER: I want to go back to something Jon said earlier about a high achieving or IB school, for example, where if the school effect -- this is a high performing school and you have all level 5 students. How are you going to show gains, student growth in those situations?

DR. DORAN: Well, remember, even if they come in really high, if we go back to the scatter plot that Jon showed you earlier today, the level 5 kids are the ones that are up here. This is the prior score, this is the current score, let's just say, right? We have that scatter plot that would look like this, but even those kids who are very high performing. Even when they come in, it's very high performing. They have an expectation and there are still students who beat that expectation. That is how we determine growth for those kids.

MS. KEARSCHNER: But with the school effect measured, I'm trying to -- I'm still not grasping the school effect in those outlined situations where you have, say, a school that's all ESE population versus a school that's all gifted or IB or something along those lines.

But in those situations does that school effect narrow? I'm trying to understand when you get to that end are you bumping up against something where it has less effect?
DR. COHEN: Point of diminishing returns. But that's my point; you get to a ceiling.
MS. KEARSCHNER: A ceiling.
MR. MURPHY: There might be a smaller range of difference between a group of teachers, but the percentages we would use would still point out the teachers that were high performers or the ones with the bigger value add. The range is going to be smaller is what you're saying, right, because of the school --
DR. DORAN: I see where you're going. There are some potential ceiling effects. One of the graphs that Jon was showing earlier there is the highest obtainable scale score and there's the lowest obtainable scale score. There are some kids who score so high that their score gets truncated to the top; they just can't score any higher. How large are the ceiling effects when we went through this; do you remember?
DR. COHEN: I don't remember and the

ceiling effects could actually be more subtle because your measurement error gets bigger at the ends of the scale. So there are ceiling effects. There is a small negative correlation -- we'll show you this later -- between the expected scores and -- yeah, we did it by -- but a small negative correlation between the expected growth and the starting point, right?
DR. DORAN: Yeah.
DR. COHEN: Higher scoring kids have a lower expectation of growth; lower scoring kids have a higher expectation of growth. The model with two lags mitigate that a little bit, you know, with two years of prior achievement, but it's still there. It doesn't go away. We're going to get into that as we start to do this.
DR. DORAN: It's a good question because, yes, there are ceiling effects and floor effects? They exist. How large are they? We'll show you some of the things that suggest there may be some issues, but that will be true with any particular measurement.
DR. COHEN: Actually, could we -- it's an important point and a good point, but it's not directly related to the school effect/no school

effect.
MS. KEARSCHNER: That's my question really. That's what I think I was trying to get at. Is there -- could there be any correlation with the school effect in those kinds of populations?
DR. DORAN: Let's take an example and answer the question. Could there be? Suppose we have a very high performing school and suppose we have a school where there are a lot of students who come in with very high scores in the intake, right? Let's just say all of the kids are clustered in this particular area. Now, can that school still show a positive effect? Well, one of the things that we saw was there is pretty significant scatter around that expectation line even at the high end of the score distribution, right? It resulted that there was virtually no scatter around that point, we would observe that students for whatever reason weren't deviating much from the conditional expectations of the high or low end of the distribution. It's plausible that being at the end of the extremes, very low end of the distribution, could limit your ability to be classified as low value-added, or being at the

very high end of the distribution if you have a cluster of kids who could be there, it's plausible. The degree to which it happens, I just don't know because we don't subset the schools when that happens and see if there is a particular consequence. But to the degree that would play out in the real world, I'm not sure is a huge concern.
DR. COHEN: This really is a big and important issue in the accountability system. You need to think about it. Unfortunately, it's not a technical statistical issue. If it was, you could ask Harold what the right answer was and he'd tell you. But statistically when you estimate the school effect, and again I'm speaking loosely but it's pretty much true; when you estimate the school effect the teacher effects are going to have the average of zero in each school. So that means that all schools would have some positive teacher effects, some negative teacher effects. The school effect will be either positive or negative. It's around that school.
Unless you believe that teachers are distributed evenly across schools, every school

has an equal mix of better and worse teachers, high value-added and low value-added teachers, you probably wouldn't want to attribute everything common in the school to something other than the average teacher effect, right? So on the one hand you want some of that to count for the teachers because you don't believe that teachers are equally distributed across all schools.

On the other hand, to attribute it all to the teacher, you may want to do that. To attribute 100% to the teacher, then you would say things like school leadership doesn't matter. If you've got great teachers in your school or you say a principal has this effect on student learning through his effect on the teachers; he or she makes the teachers effective teachers, and that's effective school leadership. So it's really a policy decision.

You can estimate the model with the school effects and then add the school effects and teacher effects together like I did on the board here and wind up with something very much like this. In that case, you'd be making the decision that I'm going to attribute everything

common in the school to the average of the teacher-effects, and that's fine. If that's how you believe it works and that's how you want to attribute it, that's fine. That's how you do it. If you agree now that you want to do that, you're probably better off with one of these models because that's done automatically and you don't even have to mess with it. But if you think that whatever is common in the school should be shifted and shared a little bit between school leadership or whoever and teachers at the school.

You estimate one of these models, but you think real hard about whether you just take this teacher effect or whether you add in all or part of the school effect for each teacher as well.

MS. FEILD: Is the school effect affected by the differentiation that we had earlier on the growth where a child who's sitting in two reading courses has a higher expectation for growth than one who is not? And the question would be on a low performing school where let's say 90% of the kids are being double-dipped versus not, is that going to negatively affect the school effect because there's a higher

growth expectation?

DR. COHEN: I understand the logic behind that. The growth expectation doesn't go up very much with additional courses and that's good to know. We didn't look at that for the schools. We did look at it for teachers, and teachers of lower performing students are somewhat more likely to have higher value-added scores.

MR. MOREHOUSE: But what if the school has a tracking system where some students are tracked in the regular courses, some are into honors courses, some are into gifted classes. Does this model take into consideration those differences?

DR. COHEN: At the margins I think it does. The teachers teaching the -- typically on average we found that even with the best of our models the teachers teaching the lowest performing students have slightly higher average value-added scores than the teachers teaching higher performing students. So a teacher who got the lower tract track courses would be at a bit of an advantage. Maybe we want that. Maybe that will attract the smartest teachers then because part of what happens with the scale,

maybe a different effect. Gina?

MS. TOVINE: You've kind of already answered that, but one of the things I was going to ask, since you've run the data for every one of those models and you have all that information, is it possible for us to see at some point like the simulation, like see teacher A and teacher B and how it would actually pan out for that individual teacher for each one of these. But you just kind of indicated how that would actually be translated over to the value-added model.

DR. DORAN: We had a slide that I think we took out and perhaps we can recreate the slide and put it back in. I think the question you're asking is, are there -- when you have the school effects and the teacher effects, do thinks change for the teachers' classification relatively large? The answer is no. One of the plots that I had previously was it showed the relationship between teachers under every single one of the models, and what you see is virtually a perfect correlation between teacher effects and models 1 and any of the models 3. What that suggests is that however teachers are classified under this particular model, the teacher only model, they remain similarly classified under this one.

If they have high value-added effects under one model, they would also have high value-added effects under the other model. We could recreate that and show that to you so you could see there's virtually a very strong linear relationship between the two.

One of the things that might have been a concern is what if the correlation was zero or even negative? Would it contain the teacher effect under this model and look at its relationship to the teacher effects under this model, and suppose the correlation became negative or zero, meaning they flip-flopped or there's no relationship? That doesn't happen. We know and we can show you that the correlation between the teacher effects under all of these models is very close to one.

Yes?

MS. BROWN: I just want to throw something out for thought, politically correct or not. But sometimes -- you know, we started our last discussion when we were here face to face

talking about wishing we could incorporate things that we don't measure, like parent involvement, homework completion, things like that that we have no data for. When we look at something like a school effect, even though the residuals are calculated based on what's in the model, we're still getting an aggregate of that school. So in some ways we're getting the hidden variable of that which we wanted to measure but we can't and we might want to think about that.

DR. DORAN: We're going to hire you because that's exactly the right way to phrase it.

MS. BROWN: Thank you.

DR. DORAN: There are things that are unknown in --

MS. BROWN: I have a very nice fee.

DR. DORAN: That's exactly why. There are things in that school, parents who do extra thinks for their kids or who don't do as much for their kids, or certain policies or practices or neighborhood effects, or peer effects or something --

DR. COHEN: There may be, but then this is really -- there's a reason that there's a

stakeholder committee involved in this process because what Harold is expressing now is opinion. He believes this to be -- he believes this to be true, the -- but the stakeholder committee here because it's your beliefs about whether those things are affecting the teacher effect. It's your beliefs about those things that should be driving the day and driving the decision.

On the other side, if you take the school effect and attribute it entirely to the school then the average teacher effect within each school is going to be approximately average, right? So there is a whole continuum of choices you could make. That choice depends on what you can believe; you're the stakeholder. You guys represent the folks who are going to be affected by this, whose children are going to be affected by this, whose staff are going to be affected by this. You need to figure out what you believe affects student learning in the school.

MS. FEILD: But speaking from the political perspective, if you're selling this model to teachers and you're saying to teachers that it's not just the math or reading teacher that

counts, but everything else that everybody else is doing is contributing to the achievement of the student, then I think you have a better chance of having a school-wide attempt to work as a team opposed to -- I mean, it's a perception.

DR. COHEN: A lot of companies decide that they'll compensate their executives partly on overall company performance and partly on the performance of their unit; so if you estimate school effects and your compensation system is compensating teachers partly on overall school performance and partly on their -- ability -- that's allowed under the current legislation then yes, you've estimated the school effects and you've made some decisions on the back end and then --

MS. FEILD: Wouldn't you say that when we continue to add additional assets to this model, when we add a APIBAICE (ph), all of that, then by having that school effect it allows you to combine all these pieces, and so the effect of all of these different assessments could contribute to a student's or teacher's performance. Is that going to happen you think?

No.

DR. COHEN: I don't know enough about all the things that are going on.

MS. KEARSCHNER: It's actually required by law that there's pieces from parents, from students. There's all these other things that have to go into the evaluation. That might be the other 50%, but it could certainly be incorporated as part of this school effect. I mean, that's another --

MS. FEILD: I was thinking when we add staying to the achievement side we eventually should be looking at models for, let's say, advance placement scores or international baccalaureate, so if those get tied into an overall school effect then again you're sort of tying in all of the achievement results.

MS. BROWN: Technically, if you tie all those different assessments together at a school level then it's all going to play into the school effects.

MS. FEILD: The teacher may be teaching English in 10th grade, AP English in 10th grade, and that student is taking the FCAT 10th grade reading test and the AP English test, right?

And in theory at some point I would assume at some point that teacher may be measured on both those assessments, both the FCAT and the advanced placement, and eventually there will be more than that.

DR. DORAN: I think this is an interesting conversation. Now, remember, the teacher effects here are essentially the unobserved thing that is happening after we control for all the stuff that we think is true, you know, the difference between kids, there's some deviation from the expectation, and there's a question mark about why that happens. In the model we assume that's because of the teacher and the school effects is kind of the same thing. There's un-observed random variations that happen at the school level that may be due to other things, and that gets called a school effect, and then teachers have to deviate from that in order to have high or low value-added. Yes?

MR. TOMEI: You talked about the consistency of the models when looked at from individual teacher outcomes, and that's using the data you created so you know that you're

dealing, I assume, or is that dealing with --

DR. DORAN: Real world data.

MR. TOMEI: Okay. Here's my question. There's been a fair amount of research done historically looking at different types of student growth models, including value-added models. The work that was done previously in Florida among the things that were done was they looked at how those models categorized teachers into quartiles, which is interesting because that's essentially what we're going to do under the legislation now; and then they looked at the stability of teacher ratings over a number of years and found way more variance than what you would logically expect to occur. At any point in time, are we going to know how these models evaluate real teachers from year to year over time and the consistency there because I think that's another important attribute that we need to consider.

DR. DORAN: I think Jon and I were the first people to write a paper on this about eight or seven years ago on the volatility of teacher effects that we observe over the time, and the paper we called it *From Saint to Sinner*

Pattern, we saw teachers that were sort of bouncing around. That's a well known effect observed in value-added models. When you estimate teacher effects on a limited subset of data, you're going to get an estimate and if you estimate it again the following year you're going to get a different estimate. There are some teachers that will switch from low quartiles to high quartiles, and part of our classification that we're going to talk about today in terms of how you take multiple teacher effects and possibly use them, it may be -- we'll show you -- when you use one teacher effect the probability of being missed by (inaudible) is very big because we don't know. There are swings.

By the term, mis-classified, what I mean is if you're classified as a high value-added, the probability that you're really low value-added could be very high. But as we add additional years of data, the probability of that misclassification goes down. We're going to actually show you data that shows exactly that in terms of what we would anticipate, and that is something that you should look at as you collect more data. We would expect more stable estimates as you use more data.

* * * * * *

(Whereupon, this concludes Day 1, Volume 1. Volume 2 continues without interruption.)

H

I

M