



SPECIAL PRESENTATION

**“ADDING VALUE TO DISCUSSIONS
ABOUT VALUE-ADDED”**

MODERATED BY:

**ROBIN CHAIT, ASSOCIATE DIRECTOR FOR TEACHER
QUALITY , CENTER FOR AMERICAN PROGRESS**

FEATURED PANELISTS:

**RAEGEN MILLER, ASSOCIATE DIRECTOR FOR EDUCATION
RESEARCH, CENTER FOR AMERICAN PROGRESS**

**CAITLIN HOLLISTER, TEACHING POLICY FELLOW, TEACH
PLUS AND BOSTON PUBLIC SCHOOLS**

**SEGUN EUBANKS, DIRECTOR OF TEACHER QUALITY,
NATIONAL EDUCATION ASSOCIATION**

**9:00 AM – 10:30 AM
THURSDAY, DECEMBER 10, 2009**

**TRANSCRIPT PROVIDED BY
DC TRANSCRIPTION – WWW.DCTMR.COM**

MS. ROBIN CHAIT: Good morning. I'm Robin Chait, associate director for teacher quality here at the Center for American Progress and I'd like to welcome all of you to what promises to be a really exciting event about value-added estimates of teacher effectiveness.

We're having this conversation at a really important time. States are gearing up to apply for the Race to the Top and other competitive grant programs stemming from the American Recovery and Reinvestment Act. We're also excited to report that the conference report for the fiscal year 2010 Labor HHS Education Appropriations Bill indicates that the Teacher Incentive Fund is going to get a significant increase in funding this year. It's going to go from \$97 million annually to \$400 million, almost quadrupling funding. This level is close to that proposed by the Obama administration and is on top of \$200 million provided by the American Recovery and Reinvestment Act. And all of these programs are going to need to incorporate measures of teacher effectiveness, so this is a timely conversation. All these programs will also need input and buy-in from teachers, principals, and other stakeholders.

In the past, some papers and discussions about value-added estimates haven't been sufficiently nuanced. There were those who thought value-added estimates had so many problems that they could never be used, and then there were those who claimed that value-added estimates represented the true value of a teacher's impact on students.

These conversations haven't always been helpful in informing policy because what the estimates mean and how and when they should be used is truly a complex issue. And for all the attention, nobody has ever stopped to ask whether the term "value-added" really fits the education policy context.

So today's discussion and paper will hopefully shed some light on how we can talk with teachers about value-added estimates and information they supply and what are their appropriate uses. How can they help align teacher policies with the goals of raising achievement and reducing achievement gaps? The paper we're releasing today, "Adding Value to Discussions about Value-Added," by Raegen Miller, offers guidance to ensure teachers' vigorous participation in conversations about using estimates of their effectiveness to inform policy. And we're lucky today to have a panel that consists entirely of current and former practitioners in education.

We'll start with the presentation of the paper by Raegen Miller. Raegen is the associate director for education research here at the center. Raegen holds a doctorate in administration, planning, and social policy, and he taught mathematics for many years in a variety of school settings.

Then we'll hear from Caitlin Hollister, a third grade teacher in Boston and a Teach Plus teaching policy fellow. The policy fellows program aims to help people advocate for change within the profession.

And we will hear from Segun Eubanks, the director of teacher quality for the National Education Association. Segun has served in various leadership roles with national, nonprofit education organizations, including as executive director of Community Teachers Institute and vice president of Recruiting New Teachers, Inc.

Then, we'll have some time for a discussion among the panelists and your questions, so let's begin.

Raegen?

MR. RAEGEN MILLER: Well, good morning. As the title of my talk and paper suggest, discussions about value-added haven't been that great to date. And there's, as Robin said, going to be in short order a lot of difficult conversations about value-added. The reason is that the stimulus has a lot of strings attached, specifically the Fiscal Stabilization Funds, which all states have taken, require that they at least enable the estimation of teachers' impact on student achievement. And then, the states that are hoping to get Race to the Top funds better have pretty ambitious plans for using these estimates.

Some states have made – cleared some hurdles. They've actually changed the law to enable the use of such estimates, but that doesn't mean that we necessarily have the tools and language we need to have these discussions, which are going to be a little bit hard. And the reasons are first of all that teachers have to be part of the conversations. There are a lot of people and they have a lot of their plates and they're pretty organized. So that makes it hard. There's also a lot of technical issues with the estimates.

Now, in the past, these tended to either get all the blame or be ignored, but the truth to the matter is somewhere in between. And then there are no estimates for a lot of teachers, over half of the teachers, and this splits the bargaining unit. That doesn't make talking easier about policies. The kinds of ways people would like to use the estimates of teachers' impact – we call these value-added – is these usually lean towards bread and butter issues. And this is scary to teachers.

And what I'm talking about today partly is that the name "value-added" is really poorly suited to the context of education. That doesn't mean the underlying estimates aren't useful. It just means that the name is poorly suited and I've heard no one to date talk about that.

So I'm going to start and I'm going to offer three tools today to help promote constructive dialogue about the estimates. The first thing to do is get a hold of a suitable name for the estimates. We'll construct one. The second thing I'm offering is a framework for relaying the seriousness of a decision – this is a property of a decision

we'd like to inform with value-added estimates – to the value-added estimates themselves and other information we have about teacher effectiveness. And the last thing is let's see if we can package all the arcane knowledge about value-added estimates into a suite of what I call due diligence principles that we can follow when we have conversations about these estimates, sort of like ships follow buoys to stay in a shipping channel, or at least did before GPS.

So we'll start with this name business and it pays to recall that the name value-added came to education from the way we estimate teachers' impact on student achievement. This is a method that's familiar to economists, especially labor economists and the name value-added makes perfect sense to them. But suffice it to say teachers weren't really consulted and policymakers and reformers kind of by default adopt the name. And now we're kind of saddled with this name that's really better suited to a manufacturing context than education.

Now, why is that? There's not really a lot written about this and I'm making an argument that has sort of two prongs. The first is that value-added alienates teachers and does so in two ways. First, accountability – test based accountability has focused us on just one dimension of teachers' practice: the ability to boost student achievement and learning as measured by achievement tests. And that's an important dimension. But there're others that matters and the name value-added kind of suggests that maybe the other ones don't matter at all or at least as much as some people think they should. And then, a lot of teachers don't have tests and there's no value-added estimates for them. And so let's say a Spanish or a PE or a music teacher might feel that they add value, yet they have no value-added in this sort of policy framework. And that seems like a problem. And it's not going to go away.

The other problem, I'd say, is that the term "value-added" is a little deceptive. It fits perfectly in manufacturing, but in education it's deceptive in that it has a suggestion of bottom-line certainty that is just not justified. There are problems with error and bias and misinterpretation with estimates. And value-added estimates are no exception. And in fact, using a name that downplays that uncertainty actually raises expectations for the properties – the statistical properties of the estimates, which is kind of a wrong direction to go.

So moving towards a suitable name, it helps to look to New York, where they, a long time ago, started thinking about rating the relative effectiveness of cardiac surgeons. And they have this term called "risk adjusted mortality rates" there. And it sounds kind of grim, but it actually is right over the plate when we're talking about cardiac surgeons, who are – the main purpose of their job is to intervene in acute situations to stave off death, so risk adjusted mortality rates speaks to kind of the core dimension of their work. It does not belittle or somehow cast aspersions on other dimensions of their work that they and others think are important, like bedside manner and the supervision of interns and medical students.

And then this part – the risk adjusted part tells us that we’re dealing with some kind of a mathematical construct, not a raw rate, and that there’s been an attempt made to adjust for differences in situations in which the surgeons find themselves. The surgeon working at the county hospital has kind of a different job than the surgeon working at a snazzy university medical center. Let’s say they’re dealing with more acutely sick patients and that’s going to affect their efficacy, how you’d measure it, and it should affect how you want to rate them relative to one another.

Now, this name has some merit and I’ll point three things out. It doesn’t slate other doctors. So an ophthalmologist shouldn’t get bent out of shape for not having risk adjustment mortality rate. It just doesn’t pertain to them. And then, it doesn’t spurn any other duties, as I’ve said. The bedside manner and other things that matter aren’t really implicated at all here. And I’ve said that it telegraphs fairness and inherent certainty in a useful way.

So I’ve used this as a model to construct a substitute for value-added to use in education. Context-adjusted achievement test effects – it is a mouthful, but it has the kind of crucial characteristics we’re looking for. Context-adjusted does this part about telegraphing, the idea that we’re going to adjust for differences in the instructional challenges that teachers face, that we know we’re dealing with some kind of estimates, some kind of mathematical construct. Achievement test helps us understand we’re talking about teachers in tested subjects and not other teachers. So the Spanish teacher, the PE teacher shouldn’t get bent out of shape. And this doesn’t seem to say anything about other dimensions of teaching that are important. It just is speaking clearly to one dimension.

Now, the mouthful does reduce conveniently to the CATES. It can be – it’s memorable. It’s generic and portable and it’s not tied to a state or a subject or a contractor or a testing company. And I think these are properties that might make it easier to start using it quickly in a wide spread way as a substitute for a value-added.

Now, I don’t know and it’s probably really naïve to imagine that we could supplant value-added in all the conversations that have to take place, but I think we should ask teachers whether we should do that or not. It’s not really up to policymakers to just plough on with the term value-added, without asking teachers. And I take some consolation in the fact that if you do a Google search, you get twice as many hits on H1N1 as you do for swine flu, which was the first name we started hearing last spring in connection with that virus. So maybe we can change names.

So the second tool I’m going to offer today is this framework that’ll help us relate the nature of a decision to the properties of the indicators of effectiveness that are in play. And it forces us to ask two questions that I think are the right place to start in these conversations. The first thing is how serious is the decision relative to other decisions that we might want to make. And then what other indicators of effectiveness do we have besides CATES and how trustworthy are they.

So here's the framework and let's just focus on the color initially. This has to do with where we locate decisions in terms of seriousness. The red region is where the most serious decisions go and the green region is for the least serious decisions. And we can make more sense of it if we try to lay some examples into this kind of spectrum of color.

So who gets a \$600 bonus? Well, I'll maintain that that's a low stakes decision. And it doesn't mean it's an easy decision to handle at a school level or that there won't be a lot of turmoil, but it is low stakes, especially compared with someone who gets an incentive to move to a high needed school. That's just more serious. And then even more serious is who gets tenure and putting that in the red zone, it sort of obviously belongs there, but even deeper in the red zone, who gets terminated for being incompetent. So that's really, really serious. There's no doubt. And I think that this tool can serve as an icebreaker in conversations about how we're going to use value-added because this was talking about where a decision falls in this spectrum is something people should know about and be able to talk about. And once you've got it there, you can move on to the question of how you're going to bring CATES to bear. And the first thing there is, well, how many other information or indicators of effectiveness do we have. All in the bottom, you'll see there's one, and two, and three. And if you're in a world where you've just got one indicator of effectiveness, then you're in trouble.

Let's say with respect to this tenure decision, the trustworthiness of CATES is just never going to be good enough to support that decision on its own. And everybody agrees about that, by the way, is just you could be forgiven for not knowing that because all the communication around this kind of thing has been essentially polluted by the politics and by bad language. So we're trying to get past that.

Now, if you have more indicators, that's better because then the average trustworthiness you need for each indicator falls. So if you have three indicators, including CATES, then the idea is that even if all three indicators are a little bit dodgy, then the chances of them all being off to mark simultaneously are pretty low. So that can support pretty serious decisions. If you have fewer indicators, you still have to maintain higher average trustworthiness.

But what this framework does for you, if you start by addressing these two questions is takes pressure off the trustworthiness of CATES because you realize that there's other things that you might going to do. Let's say you're going to use evaluations, traditional evaluations as a basis for a tenure decision. Well, if you want to fold CATES into that, then one legitimate way of alleviating concern about the trustworthiness of CATES is by increasing the trustworthiness of evaluations. And happily there's actually a lot of energy in that direction right now.

So the framework, I think, serves to help start conversations. That doesn't help us finish them necessarily, but it does produce information that's relevant to the rest of the conversation, where people feel the decision falls and also what kind of trustworthiness do we have about CATES and other indicators that should be in play here.

Now, the last tool I'm presenting is this suite of due diligence principles. The idea is we can bolster confidence in CATES once and for all by just following these principles, so make correct comparisons, use moving averages, and just use three bins. I'm going to explain each of these briefly. And I'm not going to resort to the underlying social science evidence. In fact, these principles make common sense. So they resonate with common sense. And this talk and my paper really are trying to stay out of the details of the estimation of CATES.

So the first thing about comparisons is that what the CATES do is they enable you to rank teachers and those rankings can then be brought to bear somehow on a decision. But you can rank teachers in lots of ways. And you should ask and always be very clear if you're going to advocate for the use of CATES as to what ranking scheme you have in mind because rankings could be across districts or they could be within districts. They could be across states even. They could be across grades, or they could be within grades. There's a lot of different ways to do that and I have not always heard enough clarity from advocates for using CATES in just what ranking scheme underlies the estimates that they have in mind.

So it's not hard I think in planning, unless you can make clear what ranking scheme is involved. The answers don't make sense. And I think you know you've got some kind of preposterous proposal on your hands.

The next principle is to use moving averages. And there's an analogy in baseball, right? If you have weekly batting averages of players, everybody knows, at least Americans probably all know that those averages are going to be really volatile. There's tweaks in baseball and you have to talk a much longer interval of time to expect a batting average to settle down and mean something concrete about that player's offensive production.

And the same kind of goes with CATES. We have annual estimates of teachers' impact on student achievement, but we could take a three-year rolling average of those estimates and get an indicator of their effectiveness that is much more stable. And they use this kind of thing all the time, by the way, or all over the place in Tennessee in many different ways.

Tennessee has had the technology in place to estimate CATES for a long time now actually, although they call it something different there.

And then the last principle is that let's just use three bins because once you've ranked teachers, you often want to collapse the rankings into bins. And the idea is just use as few bins as possible because the more bins there are, the more chances that there're going to be teachers falling in the wrong bins. And it turns out for most policy concerns there are really only three bins that seem to matter. The bins are which teachers are highly effective, which ones are short of that, need improvement, and which ones are chronically ineffective. And you could call these different things and you could maybe argue that there should be one more bin somehow, but really you don't need very many

bins to inform most of the kinds of policy decisions that people are talking about using CATES to inform.

So I'm going to finish by noting that I'm very optimistic that if we get quickly into lots of constructive discussions about using CATES to inform decisions, even they're low stakes decisions, which should be a lot easier to agree on, that's going to do something useful. It's going to create pressure to improve assessments. And there's a lot of energy around improving assessments now, but we're sort of starting at the standards end and we need people sort of pulling at the classroom level on the assessment end of that.

And then they also provide a lot of information because once you've got the technology in place and you're using estimates of teachers effectiveness to inform decisions about teachers, you automatically have lots of information you need to make decisions about what's going on at the school level. And we need a lot of that information too. So I'm hopeful that the tools I'm offering today will help us have some constructive dialogue that helps provide both its pressure and its information. Thank you very much. (Applause.)

Now, Caitlin is going to react to my talk.

MS. CAITLIN HOLLISTER: First, I want to thank you, Raegen, for inviting me here today. More than any other point in this report, your emphasis on teacher involvement in the evaluation design process is critical.

It seems so obvious and yet many of my colleagues are totally removed from any conversation about Race to the Top value-added measures or even performance pay. Part of that is our own responsibility, but it certainly helps to have policy leader reaching out to connect with teachers about decisions that will have a major impact on our work and our professional lives.

I want to start with a little background information about my own experience. I've never been formally evaluated in my five years of teachers. I received tenure in my fourth month on the job. And I hold professional licenses from the Commonwealth of Massachusetts in elementary and secondary education. I work in a school that's considered high performing for the City of Boston, recognized for having a strong record of academic achievement. I don't tell you this to sound impressive. Every year I have students who thrive and students who don't. I don't need to tell you that the evaluation system for teachers is broken.

Right now, tenure and license systems are laughable. I want my license to mean something and that means proving my success in some way. I'm eager for evaluation and I don't think I'm alone.

As Raegen emphasizes in his report, testing is only one tool in the evaluation process. And we're just starting out to figure out how to use it fairly and effectively. This won't always go smoothly, but it must go forward.

In order for us to figure out how to better evaluate and support good teachers to become better teachers, we're going to have to experiment with different programs to eventually create a useful process that is both fair and rigorous.

I don't want to wait for researchers and school districts to discover some perfect method. Rather I want to be part of this creation flaws and all. The framework presented in this report allows for some flexibility that's essential as we try to move forward in using CATES effectively.

When I think about teacher evaluation, I'm primarily concerned with three areas. One, I want a variety of ways to measure my effectiveness. Two, I don't want to simply measure. I want to improve. And three, when the stakes are high, I want a combination of factors to determine important decisions about compensation and tenure. I think these themes resonate as well in the report presented today.

Starting off to talk about the variety of ways to measure, here's what I have to work with so far. I look at test data, reading assessments, monthly math tests, as well as standardized test data, especially from our states, high stakes tests, the MCAS. I also have endless informal assessment data, notes from reading groups and whole class discussions, math homework, worksheets from student work during the day. I'm immersed in student data and when I take the time to analyze it closely, I can gain valuable information about my students' skills and struggles.

For example, I grade students writing projects according to a six-trait rubric, fairly standard. Then I organize the data to show me what areas of writing are particularly weak for each student and for the class as a whole. From this, I noticed that my current class struggles with word choice. So I design lessons to teach synonyms for fun, happy, sad, excited, all the words third graders love to use.

Then I can look at this writing trait after next prompt to determine what kind of progress students have demonstrated in this area. I do this alone, perhaps with a few third grade colleagues. If I were to do this as part of a more formal evaluation process, with goal setting and regular feedback, I'm confident I could be more efficient and successful in meeting my students' needs. We worry about what tests to use and how to compensate for factors between schools, as Raegen addresses in his discussion of CATES.

Just like we evaluate students differently, depending on grade and subject area, we have the opportunity to do the same with teachers. As a third grade teacher, I have MCAS data. This isn't available for kindergarten, but those teachers do have ways to evaluate their students as well. It may not be standardized, but data is available. I don't need every art teacher, every special ed teacher, every high school physics teacher to be evaluated in the same way that I am, but I do need to trust that we're all involved in a

process of evaluation and that those evaluations are tied to regular improvements, which brings me to my next point that I don't want to just measure. I want to improve.

If we dwell too long on how to measure teacher effectiveness, we lose sight of the more important questions. How do we help teachers to improve our practice? And as Raegen mentioned, the discussion of value-added and this term keeping us from moving forward has been a true obstacle. We know that the information used to evaluate teachers is incomplete. No test data, observational notes, isolated lesson, videos or analyses can provide a complete picture of a teacher's work. So we use what we have and we think about how to target the areas where a teacher needs to become more effective and more efficient.

I look at MCAS data with my third grade colleagues each year and I work with a very strong, proactive team. We analyze what questions our students are missing. If we determine that our students struggle with parallel and perpendicular lines, we figure out where and when to emphasize these concepts within our math curriculum. With a little more explicit instruction in this area, I can be fairly confident that my students will perform better on these questions this year.

In between these annual MCAS exams, I need intermediate feedback to keep track of how these specific students are doing. I can fit in lessons on parallel lines, but if my students are easily distracted or if my visual presentations are hard to see, how effective will these lessons be? I'm eager to have evaluation that combines a careful look at my test data with an equally careful look at how I apply the lessons from that data to my classroom lessons.

Analyzing my data and implementing changes on my own is quite inefficient. Working with a trained evaluator I could be making much more deliberate and focused interventions to maximize my impact on student learning.

Who do I get the most feedback from right now? My student teachers. Their questions and observations of students give me the most direct information to help me steer the direction of my classroom day to day.

This year, I have been particularly concerned with two boys in my class whose time on task is significantly lower than their peer. I don't need to collect new data to know that they are losing academic time in school. I would love to know how to sit in my classroom and design my teaching to keep them more focused, engaged, and productive.

So I went to my principal at the end of September. He's new to our school and I explained my situation. Never evaluated, still received tenure. And I asked him to observe me with these two boys in mind. We're over 60 days into the school year and while he visits my classroom several times a week, my repeated requests for advice and critical feedback have gone unanswered.

Right now, teachers like me are so used to solving our own problems or making attempts to solve them with varying degrees of success that it would be a major change in many places to make evaluation a regular and meaningful part of our job.

The problem with simply using test data alone to give teachers feedback about our performance is that it continues this pattern of teachers solving problems alone. Okay, so now that I have my test data and it might be valid and accurate and fair and interesting, if I try to apply it on my own, how much progress will I make? And here's where the framework presented today is useful.

When we start to think about how to use the test data, its place will vary, depending on the decisions being made.

So my next point about a combination of factors giving us a stronger picture, it's especially important when the stakes are high, of course in tenure and compensation situations. Anyone earning their teaching degree today is accustomed to classroom observations, even videotaped lessons and analyses, goal setting, and reflection. It makes sense to extend this as teachers assume full time responsibility for their students learning. I envy my student teachers who receive feedback from me, from supervisors, and from their peers in grad school. Then they start teaching on their own and they know they're not doing a great job, but they don't know what to do. Test data isn't going to tell them how to improve, but a skilled evaluator, whether a peer reviewer or a principal, can help those teachers interpret the data and choose areas of focus.

I mentioned earlier how I would be interested to see different systems for different types of teachers. In order to figure this out, we need to start somewhere, and I would suggest the districts to adopt pilot programs, even opt in programs that would allow teachers in a certain area to participate.

For example, take middle school science teachers, look at what data is currently collected and then add some kind of classroom based observation and analysis. These teachers would build benefit from the experimental evaluation process, while also working to adapt and refine the process to be useful and effective for all science teachers in their district or the state.

Finally, let's sell this to teachers by getting them involved in the design process. Evaluation should not be something for teachers to fear. We know that the percentage of teachers who currently receive unsatisfactory evaluations is absurdly small. Districts can afford to simply dismiss teachers who do not produce – districts cannot afford to simply dismiss teachers who produce inadequate test scores from their students. Instead, states will have to invest more in helping these students to improve in ways that are more effective than the one-size-fits-all approach to professional development.

So it's a waste of time to worry about who's getting fired right now. Teachers will buy in if they know the test data will be used fairly and purposefully. Raegen mentions three very important elements – involving teachers in all discussions of

evaluation tools, using multiple measures of effectiveness, and ensuring that the data is appropriately weighted to account for the variety in school populations and situations. If teachers know that these three elements are in place, we will begin to trust the process.

Thank you.

(Applause.)

MS. CHAIT: Thanks, Caitlin. It's really helpful to hear about how our weak evaluation system is hurting even effective teachers and the students that they teach. And it's also helpful in thinking about why policy is important and how it can really impact and systematize how we evaluate teachers and how we provide them with support and how we provide them with feedback on the practice. And now we're going to hear from Segun.

MR. SEGUN EUBANKS: Thank you. Good morning, folks. Thanks, CAP and Raegen and especially thanks Caitlin for a powerful and moving portrayal of the complexity and challenges of teaching, and thanks for giving me such a tough act to follow as well. I really appreciate that one.

I'm going to take just a few minutes to talk – give some reactions to the report, talk about what I think are some of the promises and challenges that it presents and then hopefully take as little time as I can and get involved in the more important dialogue with you. I want to acknowledge. First, I think that the report meets the objective that it talks about of creating a context for constructive conversations about these issues. Without doubt, constructive conversations are always challenging here in D.C. and providing that context is important and needed. And I think that this report does that.

I'm going to talk about a couple of the elements that I think are particularly fascinating. First, what's in the name? I must admit that there were two reactions to our CATES name change and I talked to a bunch of my colleagues throughout the building of NEA in particular and got two different reactions. The first was, "well, so what? If it walks like a duck and quacks like a duck, you can call it a duck or you can call it a cat, but it's the same thing." So admittedly that was one. Now, for some of us, myself included, had a different kind of glass-half-full perspective and an acknowledgement that language matters significantly. It matters significantly both in perception and in reality, and in how we approach problems and how we approach challenges.

I for one will never – hopefully would never purposely use the term "value-added" again and plan to see what I can do to get CATES as the lexicon. And in fact, I need to say now that if CATES shows up in an upcoming legislation on teacher quality on Capitol Hill, I want the credit for starting your revolution, Raegen – (laughter) – because I made that very specific recommendation just last week.

It matters that what we're talking about – context-adjusted achievement test effects is actually what we're measuring. We can argue whether or not that's a valid

measure, but it's clear that what we're not measuring is the value that teachers give to student lives by any stretch of the imagination. So language does matter and I think that is important that we acknowledge that.

We talk often about that. We often do things like mix the concept of student achievement and student learning. We know language matters. When we talk about student learning, it's more than just a test. I'll talk a little bit about that as well.

I think that the framework for – of the effectiveness based decisions provides a very good start for us to really talk about the value of using multiple measures about really coming down and taking perspective about the importance and relevance of a decision and how we weight it. And I think it gives us a good grounding upon which to really address issues of multiple measures and putting this context of how we use CATES in its proper perspective.

There're a couple of things that I think are pretty challenging that we need to address as it relates to this test. And one – and Caitlin certainly articulated it far better than I will, but one of the things that the report implies is that teachers don't like value-added because it unfairly evaluates our worth or worse still that teachers are quote, unquote, "risk averse." I would say many of our teachers might take particular exception to being labeled as risk averse, considering so far so many of them work in environments every day that most of us would be afraid of. And so – but while it's certainly true that our teachers don't particularly like to be measured with value-added or CATES as an evaluation instrument, the larger truth is that our teachers are simply fed up with an obsession about standardized tests that they believe harms students and harms the goals of true and authentic learning that our students need to develop.

Our teachers have told us time and again, not that we don't want to be evaluated, but that these tests are narrowing the curriculum. These tests are taking time away from everything from science and art to lunch and recess. They are destroying teachers' ability to be creative. They are destroying students' ability to develop higher order thinking skills and critical thinking skills that are so important as particularly in our lowest performing schools, the obsession with getting adequate yearly progress and making sure that we score the appropriate – the right number of our students score appropriately well on the state's standardized tests that we don't get sanctions.

Our teachers are simply fed up with that obsession and to then take a flawed system and then hold teachers accountable for it is something that our teachers challenge and feel like we need to really address. And I want to make sure that we don't forget – and Raegen talked about this concept of putting pressure on the assessment instruments. We believe that actually at the center of everything we need to do in school improvement and reform is look at what we're doing with these tests, how authentic and reliable are they, what they measure and what they don't measure, and making sure that we pay a whole lot more attention about what we value in student learning and how we measure it in other ways.

The other point that I will make is that again the report supports multiple measures, but I don't think it really goes far enough in emphasizing what those measures are and how they ought to be used. And in one way, when we talk about multiple measures, it's really two levels. The grid with the multiple measures and the importance could actually in some ways be a cube because you're really talking about two things. First, we're talking about multiple measures of measuring student learning. And again, Caitlin did a wonderful job of talking about all of the different measures for student learning that ought to be taken into consideration when we look at how well our students are doing. The test score – if the test scores are one of them, but there are many other measures for student learning that ought to create a picture of student learning. Once we've created that picture of student learning, there are then multiple measures upon which we gauge teacher effectiveness.

We don't gauge teacher effectiveness solely – even if we use multiple measures for student learning – we still don't gauge teacher effectiveness alone on the results of that student learning. We must go into Caitlin's classroom and look at her practice. We must see whether she's teaching to high standards. We must look at the skills, knowledge, and ability she brings to the classroom in real and meaningful ways. So you take the multiple measures for student learning and the multiple measures for how you measure teacher effectiveness and you develop a much richer and fuller picture for what teachers are doing and how they can be more effective in helping students to learn, both to achieve well on standardized tests and to learn the critical skills and knowledge that are important for their success.

The last thing that I'll say is just to really think about putting this in a larger context. I love the fact – and whenever we do reports in NEA, people always say, “Well, you didn't talk about this.” And so I acknowledge the fact that our reports can't talk about all of our challenges in education. But it's really important, I think, to put the challenges that we have ahead of us in this proper context.

One of the things that Raegen said on page five of the report that I found particularly fascinating – and I'll quote from that – he said, “Yet policymakers and school officials live in a world where the entire enterprise of schooling is premised on the idea that teachers actually cause students to learn.” And while I fundamentally agree with, we must acknowledge that that is in fact a brand new concept in education. Our education systems have simply not been designed with the concept that teachers cause students to learn. If you look at the history of our education system, students learn based on the way our systems are structured, one, based on their background and where they're from and the opportunities that they're given – and historically that has meant based on their race, gender, and social status. And certainly, we could argue about a couple of those, but I don't think we can argue much that the social status still is a major determinant about what students learn and what conditions that they're offered.

And second, quite frankly, the American education system is based on the concept that students learn based on their inherent ability. We track students. We see who's the smart ones and who are the okay ones and who are – they aren't so smart ones. We

divide them up and we give them resources based on that. And while many of us can argue in education today the tracking doesn't exist anymore, I still challenge any of us to go in almost any talented and gifted classroom or special ed classroom and look at the students who are in those classrooms and tell us whether or not tracking is indeed no longer existing.

So we're talking about a relatively new and powerful phenomenon, one that teachers have long acknowledged. We matter. We make a difference. We can impact student learning in real and powerful ways. Now, how can we work together – and I think that this report gives us a wonderful way to start – how can we work together, not only to – and I was going to talk a lot about the importance of doing so, not just to measure us for decisions, whether there's high stakes or low stakes, but to ensure that we improve our practice. Again, Caitlin said that far better than I could. But how can we really use systems to help us to meet our students' needs in more powerful ways and develop systems that support teachers and help them to be the most outstanding professionals they can be.

So with that, I'll allow us to move forward.

(Applause.)

MS. CHAIT: Thanks, Segun, a lot of food for thought there. So Segun addressed this issue of whether language matters. And I'm interested in Caitlin's perspective on that. Do you think that the language really makes a difference? What are teachers really objecting to in the use of value-added estimates?

MS. HOLLISTER: I think it's twofold. I do think that the language matters tremendously and at the same time I think teachers are most concerned with really how does it apply. Is this just a new term that we're going to see for a year or two or is this something that's here to stay and if so, how is it going to affect my job, my classroom, and my students?

MS. CHAIT: Great. And Segun talked about teachers being fed up with this obsession with standardized tests. So does this framework help? Does the idea of using value-added estimates within the context of other indicators of teacher effectiveness, used in combination with other indicators, used differently for different decisions, does that help?

MR. EUBANKS: Absolutely yes. I think the framework helps a great deal to help us frame the discussion. Just so long as the discussion is, again, not had in isolation of the larger picture, of challenging not only – first and foremost challenging the fact that so many of our standardized tests are nowhere near the quality that they need to be. We've done a poor job of aligning our standards and our tests for what we ought to measure and what's really important. And so if we have these conversations in tandem, I think that so long as our teachers continue to experience a huge disconnect between what they feel is important in learning and what test measures, we're going to continue to have

a problem even if we find more creative and meaningful ways to use the data. If the data itself doesn't kind grow in its legitimacy for teachers, that will continue to be a problem, I think.

MS. CHAIT: And Raegen, do you think that the quality of standardized tests matters in terms of how we use CATES or value-added estimates in informing policy decisions? Clearly the quality of tests matters in terms of engaging teachers and informing their work, but does it in fact make a difference in how we use these estimates for informing policy?

MR. MILLER: Well, the answer is the quality of the tests matters tremendously, but it's not a reason or a lack of quality of tests is not a reason to refrain from experimenting with using information derived from the tests to inform policies about teachers and about schools because there's a lot of defects in teacher practice and in the ways schools do their business. And they need addressing urgently.

And the way it boils down to me is that the tests are flawed, but they matter to students, right, and no matter how good they are, you're always going to measure student learning or student knowledge with error. There's always error in tests. And the tests matter to students, so the idea that somehow we shouldn't wrap that information up in packages that try to make sense and have it applied to teachers because teachers shouldn't be subject to that uncertainty is – seems quite unfair.

MS. CHAIT: Caitlin, what's your perspective on what information MCAS provides you if compared to other sources of information. Is it useful information?

MS. HOLLISTER: It is in the sense that it's a very rigorous test. And so if my students are doing well on MCAS, I know that they do have a certain level of proficiency. What I see as a significant drawback is it's just given one time. And so I only see that data from the class that has exited and I don't have any initial data. And what I really want and what I think my students, even at third grade, would love is to have that data in September and in January and in June.

So I think that's where we do supplement with other kinds of assessments, but again those are not nearly as fine tuned and rigorous as the MCAS.

MS. CHAIT: You don't have any interim or benchmark assessments or you've created those.

MS. HOLLISTER: We do. We do. But it doesn't match with the standardized testing. And because third grade is the first year, I don't even have previous year data from second grade tests that's equivalent. So I think that's where there's a real drawback because by the time I know my students MCAS scores, they're in fourth grade.

MS. CHAIT: Interesting. I think one of the interesting points that Raegen's paper acknowledges is that all of these measures of teacher effectiveness have flaws and

drawbacks. They all have weaknesses. For instance, principal evaluations are subjective. If evaluations use multiple raters, there is the issue of inter-rater reliability. And so each of these measures has its own weaknesses and has its own limitations. And therefore used in combination, you provide a more accurate picture of teacher effectiveness. So I think that that was an important issue raised by the report. Did you want to comment on –

MR. EUBANKS: I agree. (Laughter.) I agree, absolutely.

MS. CHAIT: – because I know that that’s an issue that you’ve talked about a lot. Okay? Great.

I’d like to open the floor to questions, beginning with the media first. Anyone here from a newspaper, periodical? Okay, then we’ll go ahead and open it to questions. Gentleman in the black.

Q: Jay Bonstingl, Center for Schools of Quality in Columbia, Maryland. First of all, I want to congratulate you folks and the center on addressing the whole issue of value-added or CATES, if you will, because it has seemed to me in research and work with school systems and doing it for the past 20-25 years that we’re really perpetrating a fraud by taking too seriously the aggregated or disaggregated data and showing non-cohort groups over time. Last year’s fourth grade test scores, this year’s fourth grade test scores is going up, going down, it really in many ways doesn’t matter because they’re two entirely different groups of kids. And everybody who’s ever been in a classroom knows that last year’s ninth graders and this year’s ninth graders are two entirely different groups of kids.

Having said that, I also want to suggest to Raegen that while it’s useful for teachers to feel empowered with the notion that they are impactful of student learning, I really seriously doubt that any teacher can, without hubris say, “I caused that student to learn.” We do not cause students to learn. We impact student learning. We influence the environment and the process and the systems in which the student learns. But the student’s motivation comes from within. It can be impacted and influenced by good teachers and that, I think, is what your work is all about. My question to you, finally, is this. The great W. Edwards Deming suggested that not everything – a statistician himself – that not everything that is important can be measured and not everything that’s measureable is important. What are those elements in the school that make students more able to come out of their school experience as good, solid human beings, as good participants in the democratic experiment, as fully engaged people in their own communities, in their own lives, in their own relationships? What are those intangibles that define measurement? And could you speak to that please?

MR. MILLER: Well, I’m pretty well equipped to speak to that. I taught for a long time. I’ve taught teachers. I’ve done my share of alienating teachers and I’m trying, in some ways, to attain for that with this paper, but the answer is not a short one. And it’s not what I wrote about. But I do draw back to your initial pointing to this and Segun also

left to this sentence about the causation. That is a kind of a symptom of me, as a researcher, being kind of obsessed with making sure that we're talking very carefully about causation. And this is not a research report. And so maybe the sentence didn't make sense. But your idea that it takes pure hubris to suggest you cause learning means that I have a lot of hubris because I'll tell you that I caused a lot of student learning when I was in the classroom, caused a lot of other things and learning is a complex process and what students brought to the classroom had a lot to do with it, as did their parents and as did my colleagues and as did – whether the electricity was on that day and all these kinds of things. So the statement probably does deserve some kind of editing at this point, if I were going to back and write it again and again.

MS. CHAIT: Would you like to respond?

MR. EUBANKS: I think that one of the real key points is that – not necessarily an argument as whether or not teachers cause learning. I agree with that. But that I don't think our systems have been set up to support that reality very well. I think that we created systems for teachers broad based that we made a pretty good deal a while back to say "we're not going to pay you compared to – well compared to what your comparable professionals are going to make, we're not going to put you in particularly good working conditions compared to someone with your skill and knowledge and what they deserve. But the tradeoff is we're going to give you relatively good job stability and a pension." And so that's the system that was created. And we're going to expect that you practice well. And what the results are is someone else's responsibility.

That was a system that was created. To now say, "well, we're going to change your system by just adding on – oh, well, never mind, we're now going to count your outcomes as well and hold you accountable without changing the surrounding peaces in that system that connected that," I think sells short how big are the challenges that we face.

MS. CHAIT: Okay, great, next question. The woman in rust.

Q: Hi, my name is Karen Smith. I'm an attorney. I represent the parents of kids with special needs. And my biggest concern about leaving the value-added term behind – I like an awful lot of what you were saying in terms of context mattering. I can look within my own family. I have a daughter who entered kindergarten reading at a second or third grade level. I have a cousin who has a child with epilepsy who – who – well, let's just say it's a very difficult task to teach her how to write her own name. It is not right to put achievement scores for those two kids up against each other and say the teacher's done something bad by not bringing them both to the same level in the same year. That would be – that's – and I think that's a big part of what you're talking about when you say context.

The problem that I see with it, though, and what I would like for you to address is how do you avoid the situation that I see a lot in my own practice, which is where you have a child who's difficult to teach, for whatever reason, a child with mild learning

disabilities that nonetheless interfere greatly with their learning in the classroom, and a tendency institutionally, too many times, for teachers to look at that or for administrators to look at that and say essentially, “your child just isn’t capable of learning.” So there’s only so much we can do. And to simply give up it’s – and I’m wondering – and I think that the detraction, from a parental point of view, of the term value-added is to say, “the only thing I care about is how much you’re teaching my child.” And that applies, frankly, at both ends of the learning scale, whether it’s my cousin sitting there and saying, “you’re not even trying to teach her how to write her name,” or whether it’s me saying, “you know what, my child came in the kindergarten reading at second or third grade level and she left reading at the same level. What have you done for her?”

So at either end – and I realize I’m looking at the extremes, but they’re really important institutionally and I’m wondering how you would address that within the term that you’re talking about and the vision that you see for use of that term.

MR. MILLER: I hadn’t thought specifically about whether – or to the extent to which parents, especially parents of students with disabilities, preferences around the term would affect the story. I was working on a pretty limited plain, trying to get us a little bit forward, to have some tools available to have conversations with the people that have seats at the bargaining table. But I think your perspective, the one you raised is an important one and it may be someone that’s grappling on a daily basis with the challenges of having really diverse classrooms, including students with disabilities, including English language learners, which – that inclusive environment does make the use of test scores for anything, a much more complex business. But maybe Caitlin would like to respond.

MS. HOLLISTER: Well, first I would say I think the more parents know about these tools and can advocate for their students in this way, the easier my job is because parents can advocate for things that I know the students need, but politically is harder for me to advocate for. So that’s essential. And I think it also speaks to the importance of adjusting some of these measurement tools to have that pre and posttest data. Of course, a student who struggles maybe comes in as an English language learner, they won’t be passing my MCAS test, but I should be able to show some growth in their reading abilities in the course of the year. And I want to be able to demonstrate that. I want to have data that shows that, as well as for my very advanced readers who come into the fourth grade reading level. I want to show that they are developing the higher level thinking skills in their reading in some way.

So I think that does speak to the importance of continuing the pressure on making these tests more useful for parents, for teachers, and for the students.

MS. CHAIT: In terms of the language, doesn’t the idea of context adjusted mean teaching students with special needs, teaching students who are English language learners, teaching students who have different – who come from different contexts. Isn’t that part of the term?

MR. MILLER: It certainly means that and that's what I had in mind. Some people might – we don't know yet how people will react to my formulation. There may be something better. Some people might interpret it as somehow embodying low expectations because this context adjustment might signal that to them. But we're going to learn hopefully a lot more about how people react to the term.

MS. CHAIT: We've been talking a lot about how context matters, but I think it's also important to acknowledge that teachers get very different results with the same students. There's a very large range in teacher effectiveness. And I think that that's what CATES accounts for or that's the value of incorporating CATES into policy decisions.

On this side, the man in the black jacket.

Q: Hello, my name is Mark Hannum. I'm an Albert Einstein fellow here in D.C. And Caitlin mentioned the idea or pilot or opt-in programs. And Raegen you've used terms like experiment. And my question really centers on what is the balance between the quick and large scale implementation of CATES systems with the goal of quickly improving classroom instruction for struggling students with taking time to develop meaningful and accurate CATES system.

MR. MILLER: Well, I think it's all about the stakes of the decisions in play. I think you can move really fast into low stakes decisions, using this information. I think it's really tenable to wait until the assessments are improved. I think you need to move forward with the assessments you have and try to use that information constructively. And for low stakes decisions, I think, let's move in a hurry. For high stakes decisions, let's be really deliberative about that and let's think hard about what other information we've got to make summative judgments about teachers' effectiveness.

MR. EUBANKS: Segun, do you want to respond?

MR. EUBANKS: I would only say that there're a whole lot of policymakers who haven't taken that good advice at hand and we've seen a whole lot of interesting policy proposals that move – that claim one to move pretty quickly. So I think that again this report and the context that we have about not standing still but moving deliberately and in the right ways is important because we've seen a whole lot of bad proposals that have basically said, "let's move full scale. Let's use CATES to make high stakes decisions. And let's do that for everybody we can. And let's do it now." And so we appreciate the perspectives that Raegen and the folks at CAP are offering on this.

MS. CHAIT: The woman up here in the tan jacket.

Q: Hi, I'm Gayle Cook (sp) from Education Policy at the Children's Defense Fund. And Segun, I think you said – I may have heard this wrong, but there're other ways of measuring teacher effectiveness than student learning? And I would like you to elaborate because I certainly can see there're other ways of measuring student learning than these tests, but.

MR. EUBANKS: Much research talks about it in three big buckets. Again, we have to remember that much of the way that we've evaluated teacher effectiveness now, well, not ideal – much of it has been – implementation, as Caitlin has talked about is not that we have bad standards for measuring teacher practice, is that we don't measure teacher practice. The principals don't come in. They're not trained. The systems haven't been designed. But we need to understand what we know about good teaching practice and expect teachers to demonstrate that in the classroom.

We think it would be – there are many ways – if you look – if you only your student is learning games, you don't ever have to visit another teacher's classroom again. You look at the results that that teacher has achieved and you assume that based on those results that that teacher is either doing well or not. And we think that that's not sustainable. It doesn't make sense for a whole lot of reasons, particularly depending on what to use for learning results. But there's a reason why we train teachers. We give them ongoing professional development. And the real goal that we have is to improve teacher practice. So the way you do that is you look at what the results are from their students, what their skill and knowledge is, what they know, what they don't know, and how they practice in the classroom. You look at all those things. You help them to improve by giving them targeted professional development and real support that they need to improve. If teachers are chronically ineffective after giving them those helps – that help – they ought to be exiting from the classroom. But if we don't use those and use only student learning, then it's just a trial and error process. We bring in everybody who wants to come in and teach it. You try it. If your students achieve, you're good. If they don't achieve, we get rid of you. And we don't think that that's a good approach.

MS. CHAIT: Caitlin, did you want to comment on the importance of using a variety of measures for assessing teacher effectiveness and what measures you think are important?

MS. HOLLISTER: Absolutely. I think – I'm relatively new into my profession and I think about how all of us when we started off, we know there're so many ways that we need to improve. And test data shows us a lot of those gaps in what our students know and what they need. And there has to be some other input to say, "okay, of all of these things that my students need, where am I going to start to focus?" And I think that's where these other pieces come in that are so important around the professional development, around even what parents want for their children, what other colleagues are working on, and why this is so important not to do in isolation. That egg crate model that we still see is far too often. And I work in a school with open classrooms, with very supportive colleagues. I know that I don't get the benefit of their expertise right now. I get a very small fraction of that. And so in order to improve that student learning piece, we have to improve the teacher learning and how we're learning from each other.

MS. CHAIT: The man in the back with the beard – Duncan.

Q: Thanks, Duncan Chaplin, Mathematica Policy. I'm just going to throw out a couple of ideas thinking about how to change the communication system. One, as you said, CATES, and I like that for a lot of reasons. I think it's really important to stress the limitations of what we've got. But I'm wondering if in the short run, sort of on your way there, you might want to think about VATS, value-added test scores. And VATS, to me, sort of sounds big. It's like oh, that's a lot. And the thing that's big about it is if you take a classroom with 20 kids and you gave each kid a test with 50 questions, you've got 1,000 data points, 50 times 20 – I think that's right. And that's a lot of information. And so that's just a thought – how you might react to that.

The second question is what do we got now instead what do we talking about merging in with it – and I'm just going to throw this idea that I came with it just now, so my apologies if it sounds silly. But I was thinking somebody's opinion without rigor – sour. So it's not that we can't do better than that, but the problem is what is it we've got in the existing data systems right now that's the alternative to that?

And then last but not least, sort of where are we going. You talk about these three bins. And that had me a little worried because it does seem to stress that value-added is the end -- that you're just going to take the value-added results and use them on their own. And my sense is that's not where you want to go and that's probably not where we want to be thinking about going. So maybe keeping the value-added test scores or CATES as a continuous measure and merging it in with other things before you make decisions with it might be something to think about. So your thoughts on those things.

MR. MILLER: Well, I'll choose two of your questions. First, I'm trying to make a pretty argument for the use of CATES. I don't like VATS. The last thing is that it might not be clear enough in the paper, but the three bins are not the end of it. And that just doesn't mean we're going to focus on just using CATES to teachers in bins. You're putting them in bins based on CATES and you have other information hopefully available that puts them in bins based on that information. And if you're put in the same bin by three different measures, then you probably a pretty good chance of believing that teacher's in the right bin. You don't know a teacher's true effectiveness, just like you don't know truly a student's level of knowledge in a certain domain that's tested, but you make estimates of that. And if you have bunches of estimates, you hopefully get them all falling – many of them falling in the same bins. So you've given me a way of maybe refining – making the paper clearer.

MS. CHAIT: Caitlin, do you want to respond?

MS. HOLLISTER: I would say that it does seem that teachers' fear and maybe rightfully so that this CATES data will be used alone and sometimes for very serious decisions. And what I worry about as well, that's a significant problem in itself, but I also don't think if we do that alone we're going to get the results that we want to see. If I start being further evaluated or compensated based on my students' test data, I don't know that that's going to help me improve my students' performance. It may give me enough incentive to work extra hours with certain students and bump up some test scores

slightly. But I don't think it's going to be effective enough. And so I think that's another compelling case for why it has to be these multiple measures.

MS. CHAIT: What about this idea of using more data? Does more years of value-added estimates help in terms of making high stakes decisions, Raegen?

MR. MILLER: You get – and there's evidence of this – you get a much more stable estimate of a teacher's effectiveness or what you're going to call effectiveness, but it reduces the number of teachers that you can expose to a decision based on that information because not all the teachers have been there for three years in sort of comparable roles. But that might be the price of assuring that you don't make really bad decision by using a super unstable one-year estimate on its own.

MS. CHAIT: The woman in a pink scarf.

Q: Thank you. I think your addition of the discussion about trustworthiness is as important and valuable as the idea of context as a teacher. But I want to ask you a question about what context means for you and specifically, does it involve measures that have to do with the amount of time that teachers have that is not obligated to student contact? I'm in Virginia. I'm in a district like Arlington, where their elementary school teachers may have at most 30 minutes of planning time, depending on their schedule and sometimes an hour. There're many parts of Virginia where elementary school teachers have no planning time, absolutely none at all. So that's my question. What's context and is time a part of that?

MR. MILLER: Well, so context is kind of a black hole, but some of those things, if you're comparing teachers or if you're ranking them across a district, all of those teachers don't have planning time, then there's some kind of fairness to the way you're adjusting for context. But in a statistical model, where you're going to estimate CATES, you can only control so many contextual values. Even if you have sort of the most robust situations, where you have lots of years of information about both the students and the teachers, as they do in Tennessee, you still have unobserved issues that affect student achievement. And those can bias the results. So you're never going to completely adjust for context. And so getting super carried away defining context now isn't going to help us, I think, use this paper for what is designed to do, which is to help people have these conversations in a hurry.

MS. CHAIT: Caitlin and Segun, what do you think are some of the key elements of context that need to be considered?

MS. HOLLISTER: Well, there are so many and it's tough. And this one that you raised was not one that I had on my list, but I think it's very valid as well. I think student population certainly in terms of students coming in with English language needs – students as English language learners, students with special needs – so the student population has to be looked at as well as the school population. I think in some ways the number of years that a teacher has been teaching we know already affects student

progress. And we don't want to be deterring teachers who are still developing their practice, but we also know we don't want students losing out because they're getting a first or second-year teacher. And so we have to be looking at that very carefully and how do we compensate for students a very novice teacher. And we know that that's significant.

MR. EUBANKS: What I'd say first, I think, Raegen really is correct about ensuring that you are doing legitimate comparisons across folks who share some of the context. And probably more important than that again there are so many context elements that you can measure that what – I think what's much more important is helping teachers to understand, right, none of us – maybe Raegen and few others in the room, my friend from Mathematica – understand the statistical modeling that is part of this CATES models.

But it's very easy to tell teachers what are the context variables we are measuring and get feedback from them. Are those the light variables in our context now and what are the variables that we're missing and how do we measure those missing variables if they're not being measured in the statistical models that we have now? Again, most models are going to use student background, characteristics and what they know about teachers in schools. But you can't put all the variables in a model no matter how sophisticated it gets. And my understanding is that the more you put, the messier it gets, or some other technical term like that. (Laughs.)

But I think really part of that is communicating with teachers and engaging them in ways that understand nothing's worse than saying you're high value-added, you're low value-added but we can't tell you, don't worry about why, it's too complicated for you. So we've got to figure out a way how to communicate what those context variables are and what's important about them.

MS. CHAIT: Clearly, most teachers aren't going to understand the statistical procedures. What do you think are the key pieces of information that teachers need to understand in order to be engaged and on board? Raegen, do you want –

MR. MILLER: Well, if I'm going to try to explain how CATES are estimated to teachers, the key thing that they need to understand is that you're working the new test scores, the students' score that you believe are somehow influenced by their presence in that classroom and then some kind of prior measures of student achievement. So this could be the score at the end of the year, before the class in question, and then sometimes there's going to be other contextual things put into the equations. But what to put and what makes sense depends on the nature of the data you've got available and that's going to vary a lot over different policy domains.

MS. CHAIT: Caitlin, do you want to comment?

MS. HOLLISTER: Yeah, I would agree. I mean, that's the central piece, absolutely.

MS. CHAIT: Great. This woman in the back.

Q: Hi, I'm Sarah Yew (ph). I'm a high school chemistry teacher, serving as a fellow at the National Science Foundation.

Raegen, you mentioned a couple of times rather emphatically that tests scores matter to students. I would argue, unfortunately, that that is not always the case. I come from California and in California the yearly standardized achievement tests do not count as part of student grades. They do not affect graduation, they do not go on the transcripts. Really, the ones – the people who are affected by the tests are the teachers and by broader extension, the schools and not the students themselves. I would contrast that with, say, India or Hong Kong, where standardized test scores mean everything to the students and affect almost everything about the students' futures.

So we've been talking a lot about how test scores should impact teachers and how much they should matter to teachers. I would turn the question a little bit and ask: How much do you think they should matter to students?

MR. MILLER: Well, they do matter to students. They matter in different degrees and depends on the particular school district, depends on the way that parents feel about test scores. It depends on the state you're in. In California – I haven't been there in a while, but I'm from there – I believe there are actually scholarships at stake based on results on state tests. In other states, that's certainly the case across the state. In some districts in California, I know local school boards have voted to put information about tests on the state tests on transcripts. I know in very high performing districts this was sort of necessary in order to help get the level of turnout on test days that was required in No Child Left Behind.

So – no, I think it's a really legitimate area of concern on how much student tests should matter for students. But it's not one I had really thought about a lot with respect to this paper.

MR. HOLLISTER: I mean, I can speak to Massachusetts even though I don't teach high school. It's certainly of great concern because these tests, the standardized tests do matter so much for graduation and we see the drop-out rate very much connected to student performance on their MCAS, which is given in 10th grade so that they have more time to pass.

And even at the third and fourth grade level, I hear students talk a lot about their text scores. So the fourth graders just got their test data from third grade probably in November this year and they're very aware of how they did. But they don't know much more beyond that. And I think, you know, when I do the assessments that I have in my classroom, I do try to talk to students about what this means, what it shows me and then what goals we can make together for what they need to do. Test-taking is a skill, and they're going to have continue to develop it. I think the more that students are able to

articulate what they do well as test-takers and what they need to practice as test-takers, the easier this is going to be for them as they move on.

And I don't know a whole lot about this, but there are some schools in Boston that are experimenting with a new data collection system that is available to the students themselves so that they can do some more analyses of their own test data and I think that's going to be helpful.

MR. EUBANKS: I don't think we have enough data (inaudible). It'll be really interesting. There's been a lot of talk about the fact that in places where there aren't actually any direct consequence to students, what does that mean about how students approach the test and how valuable it is. There's a whole lot of anecdotal data. In particular low-performing schools, students know a lot about what they're doing and how they're doing on the test and how important it is because they're communicated about it constantly on, in some cases, an almost daily basis.

And so the – in places where achievement is an issue, I think there has been some pretty significant awareness on the part of everyone in the school building, including parents, about how well their students are doing. What that means in terms of how students approach the tests and then how valid the test are as a result, I think, warrants more research, I think.

MR. CHAIT: If tests are not high stakes for students, can they be used in part to evaluate teachers, do you think?

MR. EUBANKS: That's a research question that – yeah. That's a big research question that I think researchers have talked about. I don't know how much data is available to show how – whether or not – you know, again, one of the issues – one of the things the statisticians will say is, well, if it's relatively low stakes for everybody, when that kind of takes care of that bias, or if there is a general underperformance on a test as a result of that, it's going to show – it shouldn't be significantly different across students and it should still be kind of valid for how students are doing in comparison to one another, or something like that. Am I close, Raegen? (Laughs.)

MR. MILLER: Averages mean something, whether the tests are high stakes or not and they can be used somehow to inform policy.

MR. CHAIT: The man in the black sweater.

Q: Mark Nidal (ph), independent. I have a question. It seemed that general consensus that the quality of the test is not what it could be, particularly from the perspective of teachers. You'd like to see it improved. Is the NEA or any other teacher group working independently to create a test that they would find more suitable, or are they working together and there is a political compromise issue that's just going to go on forever of fighting with other parties?

MR. EUBANKS: The NEA isn't working independently to develop testing instruments. We are working in partnership with a whole bunch of other organizations to – like – folks like FairTest and other national organizations that do comprehensive evaluation and critique those standardized tests and have experts on staff. We know and understand that developing high quality tests is possible with significantly more time and resources and so we are trying to really push for an advocacy position around taking the time and resources to develop tests that better reflect the types of learning that we think are important for students to learn, so –

Q: (Off mike.)

MR. EUBANKS: NEA has not nor are we eligible to apply for grants from the Department of Education based on our status as an association and a union. But through our NEA foundation and others who are working, again, with other partnerships to help to encourage that to happen.

MR. CHAIT: Okay. One final question. Gentleman over here.

Q: Hi, I'm Connor Williams (sp). I was a founding first grade teacher at the Achievement First Crown Heights Elementary School in inner city Brooklyn. The other day I was reading in a study called the widget effect that 61 percent of districts in the United States haven't let go a tenured teacher in the last five years – a single tenured teacher.

Meanwhile, 80 percent of principals report that they have at least one, if not more, tenured teachers in their school that they really find chronically underperforming they'd like to let go. So as someone who was placed in a charter school when was in Teach for America, what I discovered was three things that charters do that they really makes them different from public schools is that they have freedom in terms of schedule, freedom in terms of setting their own curriculum and then of course in terms of hiring and accountability.

And so there's been a lot of talk today and I'm really glad you guys have had this talk about experimenting with different sorts of tests, different sorts and schedules of observations and evaluations and what-have-you. And I wonder why don't we hold the charters accountable on this? They say they're laboratories for education policy. They're experimenting with different school models. Is there anything that we can learn, specific as opposed to saying, well, we need to experiment and improve the test and improve our structures for (conduit ?)? What have we learned? If there anything we can learn from other charter schools and what they're doing that's making them successful in this regard that maybe we're not seeing in the public school system?

MR. MILLER: Well, there is a lot we can learn from charter schools. It's not really the subject of today's talk. We don't write about it ourselves, but there are many reports about what we've learned from charter schools. It's important to note, though, that charter schools are public schools and they're taking public money and so they do

need to be included in the fold of state-testing regimes and so – and they have that information about how their teachers are doing with respect to raising student achievement and they, hopefully, can use that information and maybe because of the reasons you mentioned more robust ways to really stimulate greater growth and, you know, some of the things we’ve learned about charters indicate that that’s the case.

MS. HOLLISTER: I would say that the charters and the teachers in charter schools I know in the Boston area, I think have a lot to share with us about how they do their evaluation and peer review systems. And that’s why I do think there’s a lot of room for this kind of opt-in pilot model that I would like to see. You know, I don’t want to leave my current school to work for a charter school, but there’s some things in what they do collaboratively that are very attractive and I want that to be an option without waiting for the next contract negotiation to allow that to happen. Now, I’m a very proud union member, but I also want to see some flexibility in allowing us to experiment with some things that have been successful.

MR. EUBANKS: The only other thing I would add we have a lot to learn from charter schools. There is a lot that we’ve already learned about what – about the limitations, the scalability to large-scale reform and what happens in certain context that I think are important and are realistic to talk about. One of the things of the widget report told us in things like this are relatively clear and Caitlin’s experience with – by the way, I would not call unique by any stretch of the imagination – is that how we’ve tracked teacher performance and helped to support it is broken. The reasons for that tend to be more complex certainly than we have time to answer here.

We know there are a whole lot more complicated than the union contract won’t let us get rid of these bad teachers and there’s a whole lot of super talented, super capable people waiting outside the doors to come into our worst public schools if we could just get rid of the bad ones who are there. We know it’s a whole lot more complicated than that. And I think that, again, we’re starting to have some more substantive conversations like this and substantive conversations about how you measure and support teacher quality and teacher talent that, I think – and how you learn from charters that can be productive if we get rid of the, you know, really divisive, polarized positions that folks tend to take.

And so thanks to folks like CAP who are trying to move us into the middle, we can have these more productive conversations.

MS. CHAIT: So I’d just like to close by saying that I hope we’ve begun a new conversation about the use of CATES and how they can inform policy. And in Caitlin’s words it won’t always go smoothly, but it must go forward.

And I’d like to thank all the panelists and thank the audience for coming.

(Applause.)

(END)