THE BROOKINGS INSTITUTION

COMPARING TEACHER EVALUATION SYSTEMS

Washington, D.C.
Tuesday, April 26, 2011

PARTICIPANTS:

Introduction and Moderator:

GROVER “RUSS” WHITEHURST
Herman and George R. Brown Chair
Senior Fellow and Director, Brown Center on Education Policy
The Brookings Institution

Panelists:

STEVEN GLAZERMAN
Senior Fellow
Mathematica Policy Research

DAN GOLDHABER
Director, The Center for Education Data and Research
University of Washington

SUSANNA LOEB
Professor of Education
Director, Institute for Research on Education Policy and Practice
Stanford University

DOUGLAS O. STAIGER
John French Professor of Economics
Dartmouth College

STEPHEN W. RAUDENBUSH
Lewis-Sebring Distinguished Service Professor
Chair of the Committee on Education
University of Chicago

* * * * *

P R O C E E D I N G S

MR. WHITEHURST: I'm Russ Whitehurst. I'm director of the Brown Center here at Brookings.

Just this month the Illinois Senate unanimously passed a bill that would require new teachers to earn excellent qualifications before they get tenure. Last month the Florida governor signed a teacher quality bill that establishes merit pay for educators based on student achievement. The Arizona legislature recently passed a law that incorporates student progress data into evaluation systems for individual principals and teachers. Colorado passed a law requiring that 50 percent of teacher evaluations be determined by the academic growth of their students. Louisiana requires student growth to account for 50 percent of teacher evaluations. We recently had an event with Governor Christie of New Jersey in which he announced his plan to require teacher evaluations to be based on student achievement. In New Jersey he went on -- I'll come back to this -- proposed that every school district should design and implement its own evaluation plan.

When we turn to the federal government we've got the Teacher Incentive Fund present since No Child Left Behind. That's roughly a half-billion dollars a year available to districts to compete for and it requires that districts have to reward teachers and principals based on student achievement using fair and transparent evaluations. Race to the Top, the Obama administration initiative, required that states not have any laws in place that prevented the linking of student evaluation data to teachers and basically you couldn't get an award unless you promised to put an evaluation system in place that was strongly based on student achievement.

So you look over all of these approaches, including a lot that I didn't list just to save time, and you find that they have two things in common. One is that they require the evaluation of teachers based on the gains that their students make, and the second is that the details of how this is to be done are completely unspecified. So if we're going to leave to individual school districts in this country, of which there are about 16,000, the responsibility for designing teacher evaluation systems, we're going to have 16,000 evaluation systems. Some of them are going to be harmful. Many of them will be useless. Some of them will be okay. Some of them will be pretty good. Some of them will be great. But which is going to be which and on what basis, I don't think we're in presently a position to know.

The variation of the quality of these systems has real consequences. Teachers do not get hired in one school district and teach there for the rest of their lives. They move around. The specification of a teacher as ineffective or highly effective in one district is going to have consequences for what they're hired to do in other districts, as well as the performance in their own district. So the idea that we're going to have all these rubber rulers out there for evaluating teachers seems to this group problematic. And it becomes problematic when it's tied to efforts at the state and federal level to help districts reward their most effective teachers. So if a teacher in New Jersey has access to a significant salary increment for being highly effective, what if the teacher who is highly effective in Newark, in fact, is not very highly effective at all when you look at the quality of the evaluation system that's been put into place?

So what we've tried to do in our report is address how teacher evaluation systems that are going to differ in many operational details can be placed on the same scale in terms of their reliability. And we deal with the very vexing issue that is barely acknowledged, if acknowledged at all in the state and federal policies I've previously mentioned, and that is that most teachers teach in untested grades and subjects in which student growth cannot be measured because there are no measures of student learning.

What we're going to try to do today is talk about how you'd go about addressing the issue of the comparability of teacher evaluation systems, the problem of teachers teaching in untested grades and subjects, and the ability to have a common ruler that makes sense and can be used statewide in most states. The people who have joined me in this task are the folks on the panel. We've got Steve Glazerman, who is a senior fellow at Mathematica Policy Research; Dan Goldhaber, who is director of the Center for Education Data and Research at the University of Washington; Susanna Loeb, who is professor of education and director of the Institute for Research on Education Policy and Practice at Stanford; Steve Raudenbush, who is the Lewis-Sebring Distinguished Service Professor and chair of the Committee on Education at the University of Chicago; and Doug Staiger, who is the John French Professor of Economics at Dartmouth. It falls to them to try to explain what is a very technical, wonky report in ways that will leave you completely satisfied.

SPEAKER: Scintillated.

MR. WHITEHURST: Scintillated. We've issued, you know, previous reports. This is the third in our series, and we see that the work that we are presenting today fits in the context of those previous reports. So we're going to be covering previous material, material that's released today, and also trying to put this in the broader context of improving schools in general.

I'd like to turn to Steve Glazerman to provide a brief review of the material in our previous report. It focused on value-added and what's good about it, and that's an important context for the rest of our work. So, Steve.

MR. GLAZERMAN: Thanks, Russ.

So the report that we are discussing today is the third in a trilogy. One of those earlier reports released last November was about the importance of value-added measures in the overall goal of measuring teacher performance. And I guess I should probably start by just saying when we talk about value-added we mean teachers' contribution to student achievement growth over some specific time period. So when we talk about value-added measures -- that is our attempt to measure that concept -- then we're talking about using student test score data and data on the characteristics of the students that each teacher has and their prior achievement to control for as many of the factors that we can that are outside the control of the teacher.

So the paper that we released in November was really trying to address some of the concerns that have been raised about value-added measures and to put them into context of a decision -- policy decision-making process. And the basic conclusion is or argument that we tried to make is that value-added measures are not perfect and they're -- many of the reliability concerns that many people have raised once value-added as an idea has gotten a lot of attention -- thank you to the *Los Angeles Times* and the *New York Times*. We're not disputing any of the issues about reliability but what I think we need to sort of work through is how to use -- how to incorporate the information on the reliability of a teacher performance measure like value-added into the decision-making process so that the policies and the stakes attached to those policies are aligned well with the quality of the measures themselves.

And we made a number of specific -- really four specific arguments to sort of guide policymakers through this issue. The first one was really about value-added measurement versus accountability. Okay? So oftentimes there is a lot of debate and controversy around value-added where much of that controversy is really around the particular policies that they're used to inform. And so we want to just point out the distinction that value-added measurement itself is different from what you then go do with it. It's important to, you know, to very carefully align these policies with the quality of the measures themselves and not overreach but at the same time understand that the measurement itself has any variety of uses which can include summative uses, human resource personnel decisions, as well as basic issues like, you know, more formative policies such as professional development, targeted professional development, or teaching assignments, helping teachers understand their own strengths and weaknesses.

The second point that we made in this paper was about the consequences of misclassifying teachers. Even in the absence of value-added, classification decisions are being made all the time. So in sort of a default teacher evaluation system or a very common teacher evaluation system, it may be the case that 90 percent of teachers are given the top rating in the system or 95 percent or 98 percent. In that case, if the teacher evaluation has no bearing on, let's say, a tenure decision, tenure is automatic, that policy will have a very low probability of denying somebody tenure who deserves it. And that's a good thing. But it will also have a probability of granting somebody tenure who might not be effective in the classroom, and even with intensive professional development or other kinds of assistance might not be successful.

And so the point that we made in that paper was that each of these tradeoffs between what we call a false positive, you know, identifying somebody who is labeled as exceptionally weak who isn't, is -- we can sort of reduce the number of those false positives but at the same time we might have to increase the number of false negatives which is failing to identify somebody who isn't. And the key idea here is that the consequences of false positives and false negatives within this context -- and this exercise can be repeated for any decision, whether it's extra compensation or targeted professional development -- is different for teachers than it is for students because the consequences for a teacher, you know, more job security, sort of less stigma, those are all positives associated with one system. But if you go very far in that direction then the students -- you increase the likelihood that a student will have a very weak teacher. That's not identified as such. And so we felt it important to recognize that tradeoff and for policymakers to be real clear on what those -- that they are making those tradeoffs when deciding whether to use value-added information and to what degree and how much weight to place on it.

And the third point -- and actually the third and fourth points are sort of related so I'll combine them -- is really we could frame it as a question which is how reliable do value-added measures have to be in order to be useful? And the really surprising, somewhat counterintuitive answer is not necessarily very reliable. Where do we sort of come up with this? Well, if you look at other professions, you know, highly complex professions, like teaching, and you could say policy research, education research, or professions like medicine where outcomes are hard to measure, that the job tasks are very complicated, we often rely on very -- on sort of basic measures that aren't necessarily highly predictive of future outcomes but they're somewhat predictive.

And to give you an example, actually, we drew examples from a number of professions but the one that sort of caught a lot of people's attention was professional baseball. Now, a batting average is surprisingly not that predictive of future batting average for a professional baseball player. About .36. So the correlation between batting average this year and next year would be about .36. When you look at the correlation between value-added measures for teachers in a given year and the next, the estimates vary quite a bit depending on the context. But sort of looking across a number of studies they can be anywhere between .1 and .5. Most of the estimates are between .2 and .4 and, you know, the .3, .35 range is right around the same place as professional baseball players -- predicting future batting average. And many other complex professions where some of this research has been done we found -- researchers have found similar predictive power. SAT scores, for example, predicting freshman GPA, again, the combined verbal and math freshman GPA -- sorry, combined math and verbal SAT predicting GPA with a correlation of .35.

And so the question is what do we do with this information? If you look at how others use this information, typically it's one part of many other factors that are used to determine college admission. Many other factors to determine whether to put this baseball player in your lineup, many other factors to determine where a teacher should be placed, what kinds of professional development they should have access to and so on. And so an important point that we make in our paper is that -- well, I guess two points -- is that there's a lot of noise but there's also a lot of signal contained in value-added measures. And if we ignore the signal altogether, we really miss out on important information.

And we provide an example of what would we do in the absence of value-added information. If you look at teacher layoffs, usually the predominant determinant of whether teachers are laid off is experience, years of experience. And if you look at the average value-added or the average predictive student performance when you lay off teachers just from the bottom end of the experience profile versus the bottom end of the value-added distribution, you would get a very different result. And so in some cases it could be fairly dramatic what you can do by just incorporating an additional component into a teacher evaluation system.

And so the goal of this was really to, I think, sort of wrapping it all up, is to understand that value-added has an important role to play. The value-added indicators themselves provide a substantial amount of data but we should never take our eye off of the limitations and ways to improve these measures. And one of the things that we began thinking about as a result of this effort I think was different school districts and states are going to be in different places in terms of how much data they have, the quality of the data, and the size of their pools with which they can, you know, make statistical adjustments and the reliability with which they can measure teacher performance. And we thought what kind of relationship does this have in terms of how much weight should be placed on these evaluation systems and how useful they can be for policymaking? And that's what led to this -- to the discussions that we're going to get to the rest of this panel. Thank you.

MR. WHITEHURST: Thank you. The report that's released today is built around a particular case study and that case study is a proposal that this group put forward about 18 months ago to create a federal system for recognizing superior teachers and giving them a portable credential. And, of course, the ability to do that requires some confidence about the reliability of a system that would be used at a district or a state level to identify those teachers. And so Dan Goldhaber is going to give us a little bit of a refresher course on that proposal because understanding it will allow us to see where the reliability and quality issue as it goes to the issue of how good the evaluation system is, how that fits into practical issues.

MR. GOLDHABER: So I'm actually going to back up and say that when

Russ convened the teacher quality task force we began by kind of thinking about what are the big issues in education vis-à-vis teacher quality and what might -- how might we as a bunch of researchers contribute to those issues?

And I think that we kind of centered on two big issues. One is that there's inequity in the distribution of teachers across students, and the other is that we know there's variation. We know this sort of anecdotally and now through a lot of empirical evidence in the effectiveness of teachers, but as Steve pointed out, teachers based on performance evaluations all look like they're excellent. So the document, the track record for teachers does not reflect the reality. And if the documented track record doesn't reflect reality then it's very difficult to act on the differences, you know, that do actually exist among teachers.

So coming out of those sort of two big notions was the America's Teacher Corps or ATC. And the idea of the ATC is that teachers who were in states and localities that had reliable teacher evaluation systems -- concentrate on the word reliable because you're going to hear it again. I'm looking down the line. Reliable teacher evaluation systems that give you a verified distribution of teachers. So no widget report. Not everybody is excellent. But you actually have to have -- you have to be able to place teachers on a distribution.

So if you're in a locality that looks like that and if you're a teacher who's in the top quartile of the distribution for three years, you can become eligible to be an ATC teacher. And what being an ATC teacher means for you is it buys you a portable credential so you can work across different state boundaries, so that's quite different than the current teacher licensure system, so long as you're working in a high poverty Title I school.

The value on top of the portability is that it’s a means for the federal government to target some money to disadvantaged schools. So the federal government would supplement teacher salaries. We costed out the program if the salary supplement was $10,000 per year. So federal government targets teachers, high poverty schools based on a reliable teacher evaluation system. In a nutshell, that’s the program. Except that I should mention that we sort of mandate that value-added ought to be a component of it if you’re in a grade and subject that’s covered by value-added. But it ought not necessarily be the sole component that determines the evaluation system.

I want to spend just a second talking about the political calculus. And I’m glancing over towards Michelle. I’ve got plenty of time, right?

All right. So the thing that I think is really kind of useful about this is some of the political calculus because there’s always questions about the federal role and federal overreach, and that’s probably particularly true now. The nice thing about this is it doesn’t prescribe how states and localities ought to be judging teacher effectiveness per se. There are lots of different systems that they could come up with so long as they are reliable systems that create, you know, the spread of performance evaluations. But it gives the federal government a means of targeting money to high poverty schools, and I think that it gives teachers a real impetus to request of their employers that they develop good evaluation systems. So I think that we see the program as having an important but narrow role but could have lots of beneficial ancillary effects because if states and localities develop good systems then there are a lot of things we could learn about teacher effectiveness, teaching, teaching context, et cetera.

So that is in a nutshell the ATC proposal. And I’m going to turn it over to Doug who gets the difficult task of beginning to describe what it is that we do in this report.

MR. STAIGER: So I get the wonky part. And so I think the key is not the technical issues which you can get out of the report. And obviously I'm not going to be talking about those right now. The key is the kind of big concepts that are in here that I think go beyond -- ATC is a useful example to work through some of the issues but they go beyond that to any time we're trying to identify exceptional teachers. But that could be at either end, either the ones who are most effective or least effective for multiple purposes.

So the key message in this proposal we have for evaluating teacher evaluation systems is the way you evaluate -- that we're proposing to evaluate teacher evaluation systems is based on their ability to predict future teacher performance on an agreed upon yardstick. So that's both the idea that this should predict something that's persistent about the teacher that's going to stay with them and we're going to have to agree on a yardstick -- what it is we're trying to achieve. So we'll come back to that.

And then within that framework, a stronger association of the evaluation, however we're doing our evaluation. It may take multiple components and put them together. It may be a principal's best judgment. Whatever we're doing. But a stronger association of this evaluation with future performance is going to allow the system to identify more exceptional teachers. So it's going to be kind of a sliding scale. The better your evaluation system is, the more teachers we can identify for this exceptional status, whatever it is. And those are the two things at the heart of kind of the proposal that we're putting out.

So let me start. The first thing you have to do here is decide, you know, make your policy decision. What's our aim? So let me -- in the ATC, we made a really specific aim. We said we want to identify and reward teachers whose impact on student achievement places them in the top 25 percent of teachers in their state or district. Top quarter. So there are two parts of that. We decided how exceptional, top 25 percent. And we decided what's the yardstick, and for us it was value-added. It doesn't have to be. Right? That could be many other things. So you've got to choose the common yardstick. We chose student achievement as measured by value-added on a state test, but it could have been -- in other situations people might say, no, we're going to do student achievement on a different test or something -- not a state but some other test or it could be that we're going to look at other student outcomes as measured by, you know, some student survey. There are a wide range of things and that's a political decision. The key is once you have that and once you have this level of exceptionality, what are we looking for? The top 25 percent? The bottom 5 percent? Those are policy decisions. Once you have those, then this approach you can apply. So what's the approach?

So given this aim that we have, how do we determine if an evaluation system passes muster? And our proposal really has three important features. The first thing is our approach is based on this relative strength of prediction. So how reliably does the district's teacher evaluation system predict future teacher performance? And the reason for that is there are many things that happen in one year that can't be replicated, that aren't there, that the teacher cannot do the next year. They happen to have a very good year this year but that doesn't mean they were a good teacher. They had a good class or had, you know, one year they have a disruptive child or they had a bad year outside of school. What we're interested in for most of these, and certainly for ATC, is identifying this persistent part. Okay? And sometimes we'll refer to that as true teacher performance but what that means is persistent.

The second component of this proposal, important feature, is that the school districts with more reliable evaluation systems can identify a greater proportion of their teachers as exceptional. So it’s a measured response. Not an all or nothing. It’s not like your system has to achieve some minimum threshold and once it does, okay, 25 percent of your teachers are eligible. Basically, the better your data, the better your evaluation, the more teachers we’re going to be confident in saying are truly exceptional.

And the third piece of this is this approach -- the approach that we proposed, and there’s some technical detail in the paper, minimizes mistakes, both over inclusion and under inclusion of teachers in the program. Teachers who get it who shouldn’t, who aren’t actually in the top quarter; people who are excluded who were actually in the top quarter. And it’s going to maximize the average difference in the benchmark, the yardstick in value-added between the people who we identify for ATC and those we don’t. So it both minimizes mistakes and maximizes the difference in performance. Okay? And that’s the -- those are the three pieces that are kind of key to this.

So how do we do this? That’s kind of the big picture. So the key to this is -- the first step is you’ve got to determine the correlation, the association, the strength of the association between the overall teacher evaluation, however the district comes up with it. It may be a very formal process, it may be informal, but at the end they have this number, 1 to 100, whatever it is. And so the first thing we’ve got to determine is the correlation between that overall evaluation and the teacher’s future performance on the yardstick, which for us is value-added.

So there’s a technical detail. The correlation has to account for the fact that the future value-added is measured unreliably. So even if we knew the teacher’s true persistent it wouldn't be perfectly correlated with it. Okay, technical detail. Sweep that under. Suppose you can -- there's a way to adjust for that. And basically it's saying we're not interested in the non-persistent stuff, just the persistent impact teacher has on what we care about. The overall evaluation that we're going to be looking at, if we combine non-value-added measures and value-added when it's available, the non-value-added measures could be things like, you know, observation in the classroom, student-parent surveys, anything you have, principal-peer evaluations, et cetera.

And then a key issue is most teachers don't have value-added. There are all kinds of non-tested grades and subjects. And unless we're going to really go whole hog on testing that's going to always be the case. So for those teachers we're going to need to know something about, well, if we evaluate them just on the non-tested, the non-value-added measures, how good is that as a measure of what we care about, which is something we don't observe in the value-added? The way we're going to do that is we're going to say for teachers who have both, if we just evaluated them on non-value-added measures, on their classroom performances, how good would that predict those teachers value-added in the next year? So in some sense we have to make it a case that the amount that performance in the classroom predicts value-added in math tells us something about how performance in the classroom in history would have predicted their value-added had we been able to measure it. Okay?

So now we've got this correlation that we've measured for everybody. What does it do for us? So the strength of this correlation determines the percent who are going to be eligible for the status, exceptional status. For us, ATC. Think of the correlation if you square it is the percent of the variation and future value-added that can be explained by the evaluation score. Higher is better. You know, if you can explain 100 percent it means I just nailed it. I know your future performance. If it’s zero it means you were throwing darts. So that’s the issue.

So perfect correlation, that means the people who were ranked at the very top, at the top 25 percent on your evaluation are exactly the teachers who are the top 25 percent in the future on value-added. So if you have a perfect correlation, I just take the top 25 percent. They’re all eligible. I’m sure they’re all in the top 25 percent. If you have a zero correlation, that just meant I randomly chose people and made them very top in my evaluation and randomly chose other people and made them the bottom. No difference in average performance in the future. And in fact, in every group about a quarter of the teachers who we identified as the very top and the very bottom are in fact the best, in the best quarter. Right? Because we were drawing randomly. So that’s no help at all. It didn’t distinguish at all so we don’t identify in that case any teachers for ATC. Your evaluation system is so bad it provides no information. We just won’t flag any teacher because we don’t -- we’re not sure.

The intermediate case? The argument that we make is that you should include -- and this is a parameter that you can choose differently but we chose a particular, this tolerance parameter -- the question is we said you should include a teacher if they have at least a 50 percent chance. So the weight of the evidence says that they’re in fact in the top quarter of effectiveness among teachers. And think of, you know, so when I know for sure, when I can predict perfectly their future effectiveness, I know for sure the top 25 percent are really there. But in this intermediate case you can’t always -- a lot of people you aren’t sure. So in this intermediate case, the stronger the correlation, the more the teachers will achieve this threshold of the weight of the evidence says I think they’re in the top quartile. So when you have a very low correlation you might only pick the top 1 or 2 percent because I'm pretty sure they're in the top. As that correlation gets stronger and stronger, I start getting a bigger fraction of teachers who I think pass this threshold.

Think of this. This is a weight of the evidence standard, like in civil cases. It's not a beyond a reasonable doubt standard, like in criminal cases. What this does for us is it gives you -- it will minimize mistakes if you put equal weight both ways on mistakes, making mistakes of omission and commission, and it will maximize the difference in value-added between the two groups.

So, you know, even -- by the way, even in systems that have -- that use value-added as part of the evaluation, they won't perfectly predict future value-added because value-added has so much noise. So even in a system with value-added you might do better by adding lots of other things into your evaluation because it will give you a better indication of who's going to be most effective in the future because value-added measures are themselves not very good. You'll have lots of noise in them.

So let me give you a quick example and then wrap. So the quick example we have in the paper based on evidence from some teacher evaluations and value-added taken from various studies, the evaluation score that used kind of everything explained about 50 percent of the variation in future value-added. This was kind of -- you could think this was related to a Florida example that had principal and value-added. Principal ratings and value-added. So it could explain 50 percent of future value-added for teachers if I use everything, but if I only use the non-value-added, so for a bunch of teachers, I can only explain 10 percent of their future performance. So there's a big loss by not including value-added but you can explain something with principal ratings.

In those ones where I observed value-added and make the complete score, combine everything, the top 18 percent of teachers would be eligible for ATC. That many teachers, we're 50 percent sure, the weight of the evidence says they actually are in the top quarter. For the folks who we don't have value-added, only 2 percent of those teachers are we really sure. So you really pay a cost as the system -- as the quality of the evaluation system goes down.

The difference in value-added between the folks who we flag in and out of ATC using this method, it's about three to five weeks. Five weeks for the people where we have value-added, the best evaluations, and about three weeks for the people where we don't.

SPEAKER: Of student growth, right?

MR. STAIGER: Huh?

SPEAKER: Of student growth?

MR. STAIGER: Of student growth. Three to five weeks.

So that's the system. It's kind of figure out how good, you know, evaluate the system based on how well you can predict future teacher performance on this yardstick and the stronger the association of the evaluation with future performance means the more people you can identify as exceptional because you've become more sure of those decisions.

MR. WHITEHURST: And so Doug has highlighted in the example the problem with the system in Florida that uses only principal evaluations and accounts for very little of future growth. The nature of the additional components of the evaluation system are critical and Susanna is going to talk to us about that.

MS. LOEB: Okay. So our goal was to look at -- we have different districts or states -- in this case, different districts with different evaluation systems and we're going to see which system is stronger or weaker depending on how well they do at predicting future value-added. So value-added is an important part of the system that we have here. And there are a couple of reasons to think that value-added is good and the reason that we chose it for this example.

And I think the first reason is that it is a measure of student learning and we care about student learning. I think the real reason, the second reason that we chose it is because all the policies that Russ talked about shows it and we were just starting with something. But it doesn't necessarily -- for the models that we have here it's not essential that it's value-added. That's just -- we chose it because that's what the state has chosen. And what my role here is to talk about, kind of the strength and weaknesses of that and how -- what alternatives there are. So value-added on state tests again has the benefit of that's what these evaluation systems are supposed to use and that it's based on student learning on something that the state has agreed is the goal of schools. So those are important things but it does have a few shortcomings.

One of the shortcomings which is clear is that it's usually only available for a very small subset of teachers. Okay, so you've got only a few teachers who have it and many teachers don't and you don't want a system that only evaluates some teachers and not others.

A second one is that the current measures, even for teachers who have value-added are kind of narrow in the set of domains that they cover. So they cover -- for an elementary school they'll cover math and English, language arts, but they won't cover history and they won't cover science or motivation or attendance or, you know, your engagement in school or things like that. So there are reasons that we would want different outcomes as well.

Another issue which we've talked about a little bit here is that there's error in attributing the value-added -- well, there's just error in this measure of value-added. So maybe something happens to a child during the year and that's a shock. It's difficult in the models to adjust for that and so things are lower this year than another year. Or maybe there's a positive -- positive things that happen in one year. Something comes out that really -- or something happens that really helps the kids do better and so you have a positive shock. So there's this measurement error in it that makes the value-added not a perfect measure for each teacher. So that's a third one.

There's a possible bias in it. So we try our best to attribute things to the teacher in terms of how much kids learn but it isn't always the teacher that determines that. Something else may be happening as well. I guess the negative shock to a child would be one indication of this.

And then finally, aside from just as an evaluation tool, a negative or a less positive part about value-added is it doesn't provide a lot of information about how to improve. Okay, so there may be things in the way that you want to evaluate teachers that value-added can't do for you.

Some of these things are important for kind of the system overall, and some of them are really important for the model that we're presenting here. So, for example, the idea that it doesn't apply to all teachers, that's really central to what we're talking about here. What do you do about that part of value-added not applying to all teachers and how can we adjust for that? And the second part that's really relevant to what we're doing here is the measurement error, the fact that sometimes -- it's not perfect. Something might happen on test day that makes your kids look lower than they should look or something like that where it doesn't really -- it's not a good signal of how good you are as a teacher.

So we want in our system to try to get around those two parts. There are reasons that you would want to get around these other things, like potential bias and the value-added being too narrow or that you want to provide information. Those are important things to consider but not the ones that we're necessarily focusing on here.

But in either case, whether it's the ones we're focusing on or not, you need alternative measures to supplement the value-added measure from the state. Let's say we're just talking about the case where not all teachers have these value-added measures. There are a whole bunch of different approaches that you can take to the alternative measures. I like to classify them in two different ways or in two different kinds of approaches. One approach is to collect alternative measures of student outcomes. So you could collect information on student motivation and how that changes over time or they're learning in science or they're learning in history or something like that. That's one possible approach to an alternative type of measure for evaluation.

But then there are other ones that aren't based on student learning and what's happening to them but based more on outside assessments. So just the more subjective assessment of a school leader. It could be a very structured observational protocol by a professional who comes in and observes what a teacher is doing during the day. It could be surveys of parents and students. There are a whole range of things that you could use there.

And for the purpose of what we're doing here, we want to think very specifically about two things those alternative measures can do. And one of them, well, one of them is we want to be able to -- I guess separate from these two things -- but we want them to be able to be done for all teachers. And so many of these things are much easier to collect for other teachers. But I think there are two things to think about. One is that we want these measures to improve our prediction of value-added in the next year. Okay, to make that prediction better if we possibly can. And the second thing that we wanted to do is we want to measure effectiveness for teachers that we don’t have value-added for.

Okay, so that’s what we’re looking for these alternative measures to do. To improve the prediction of value-added in the following year and also to measure the effectiveness for teachers who don’t have value-added. So that’s when you’re thinking about it in this context. When you’re thinking -- I guess another way to think about it is that for our purpose we want to use it to measure things about other teachers and to deal with the measurement error. So the two things we’re dealing with are measurement error and not having it for other teachers.

But there may be a lot of reasons that we would want it for improving the system overall, and in particular we may want it for improving the yardstick that we use. So, for example, we use this yardstick of value-added for -- based on state tests, but the state may determine that that is not that there may be some bias in it, it doesn’t include other things that we want, so the state could say I want a yardstick that includes more. That includes an observational evaluation. That includes something about kids’ motivation. That includes something else. I think we can use these alternative measures for that and that’s a really important thing for us to think about in the future, similar to just improving the tests. That’s not as relevant for what we’re talking about here and this is -- I’m kind of here, I think, to narrow the focus down. To say, okay, for here we want these alternative measures because value-added has measurement error and because we don’t have it for all the teachers.

MR. WHITEHURST: Good, thank you. All of this occurs in a broader context of school reform. We don't get to think about that much because this is -- we have been very much down in the weeds on how to evaluate evaluation systems. But we need to look up at the horizon for a moment and think about how it fits into the large agenda. And Steve Raudenbush is going to help us do that.

MR. RAUDENBUSH: Russ, do I get that honor because I have more gray hair than anyone else sitting up here? But I'm glad to have that opportunity.

If we look broadly at the teaching profession in our country, four things seem to stand out. The first is -- I'll call it the privacy of practice. Teachers work quite autonomously behind closed doors with little guidance devising their own ways to teach mathematics, science, history, et cetera. Think about the average level of preparation of an elementary school teacher in mathematics. Give that teacher a modern textbook which tries to teach a modern approach to set theory to little kids. Put that teacher behind a closed door and say go figure out how to do it. It's scary. Okay? So that's the first thing you'll probably see in practice.

The second is the vast variability in teacher effectiveness. And Doug Staiger has done a lot to help us establish the scientific basis for this. This used to be something we would debate. Do teachers really differ? Do they really matter? Is it all student background? We have very strong scientific evidence from randomized experiments showing that teachers are dramatically different in their effectiveness. Which shows us the power of instruction but also shows us that we have intolerably great variation, particularly the bottom end, which I think helps us explain why we don't do well in comparisons with other nations. That's more speculative. So vast variability. These things are connected. The fact that teachers work autonomously in the privacy of their classroom with little guidance is connected to the fact that there’s vast variability in outcomes.

The third fundamental characteristic of the teaching profession is the utter predictability of teacher compensation. Go into any district and tell me two things. What is a teacher’s highest degree and how long have they been in the system. And we can predict virtually with certainty their salary. But yet we know that those two things are almost completely uncorrelated with this vast variability and expertise. Okay? So we’re giving highly differentiated rewards that have no relationship to how effective people are.

And then the fourth fundamental feature is that most teachers are working with lack of information. In fact, the whole system lacks information at every level. So here’s the way I like to think about it. Teacher gets up in the morning and goes to work and thinks have I been effective? And basically you judge it by how well your kids seem to respond. You know, they respond to some extent but to some extent they didn’t do as well as you were hoping. Well, you have an alternative theory. Maybe it’s because the kids are from -- they’re having a tough time at home. Maybe they’re from a bad neighborhood. Maybe because English isn’t their first language or they have a disability. So you always have these two hypotheses about why the kids aren’t doing as well. Is it me or is it them? And teachers talk about these things.

The only way a teacher could ever resolve that question would be to know how other teachers are doing who are teaching similar children. It’s the only logical way I can actually think of of resolving that contradiction. Which is it? If I see that other teachers who have very similar children are doing extraordinarily well, I have to then reject the hypothesis that it’s something wrong with the kids, their neighborhoods, or their families. Right? So that’s got to be really fundamental. And I think that’s a fundamental feature of this reform, is to give -- is for each teacher to have information about how well they're doing relative to how well they might do with similar children. That's really kind of at the bottom of all of this. Ideally, of course, not only the teachers but the principals and district officials would have information of the same kind. Right now the district officials basically know what fraction of the kids in the school are proficient. Beyond that they really don't know very much.

So you add lack of information to these other three characteristics and, you know, we have a big problem. These things are all interrelated because it takes information to improve practice. It takes information to reduce variability in practice, et cetera. All of these things are connected. But information plays a key role.

So today's event is an attempt, and the three studies of which this is the third can be regarded as an attempt, to open the classroom door just a crack, take a peek in, gain very partial information that can be used for a very limited purpose, namely to recognize those teachers who are really making magnificent contributions to student learning. And I think that's a very good thing. But we should also ask what would our schooling system look like if we opened the classroom door to the bright light of day and if teachers had much better information about how well their kids are doing and about their own level of expertise? If everyone understood that other people were more or less expert. If the most expert people had an incentive to help the least expert gain expertise, if schools were organized around this information, you know, what would our system look like? And I have to say that I'm writing a book about this. And I'm not -- you'll be very happy to know that I can't talk about it because I'm limited in time because if you get me going I wouldn't stop.

But I can say that some very terrific schools that we see in very disadvantaged areas, in those schools teachers frequently assess their kids in a way that the assessment system can't be gamed that all of the teachers not only know how well their kids are doing, they know how well other kids in the school are doing and they all know how well other teachers are doing. But in that setting, it's not a terrible thing to lack expertise. Novice teachers tend to lack expertise. The idea is to motivate people to want to gain expertise, to leverage the more expert teachers who then help the least expert and reward them for it. Give them compensation for taking that kind of leadership so that you end up with a system -- an open, more public system in which there's information in which there's a lot of guidance and a reduction in heterogeneity of the teaching skill. And the whole idea is the entire school is accountable.

Now, that's kind of -- I realize it's kind of a utopian perhaps vision. But I think it's good to have ideas. It's good to have goals. And I think that what's exciting about this event and this series of papers is that by opening the door a crack and looking inside, it focuses on us and a number of serious questions if we take this report very seriously.

The first one is one that my colleagues have been referring to, which is focusing on what should be the benchmark, the yardstick as Susanna said. What -- we had a lunch discussion with some people earlier. Some people think it shouldn't be state standardized test scores. It should be something else -- a richer representation of what kinds of gains we want kids to make. I tend to be -- I tend to think that part of the formula should involve attendance. I work, you know, in an urban school system where getting kids to go to school is a crucial part of what's going to get them to be successful. So having people be accountable for kids showing up and being engaged as well as their test scores. So we can have that debate. That's a very healthy debate and that's a good thing about this report. It really puts a laser focus on that. What should be the benchmark?

The second question then that follows logically is, well, what are those other measures? What are those classroom observation measures? What is that expertise that the principal requires in order to bring more information to bear? And the key point is it has to predict the benchmark. Do you see? It's not just because we think it's a good idea; there's an empirical test. But we can focus discussion on that. The Measurement of Effective Teaching Study funded by the Gates Foundation is very systematically looking at a number of very interesting and potentially important systems of data collection asking kids about their experience, observing classrooms that we think will have potential to predict children's learning. So we have this richer array of things that's focusing attention on that.

The third question then is how would we then use these -- this data, i.e., the benchmark itself but also the other information. In an evaluation? To make decisions? To minimize errors, errors of either proclaiming a teacher who is not so good, good; or a teacher who is good, not good? And how do we actually do that? And I think that this report gives you a very clear way of optimizing on that question. We're not going to get it perfect. The best system is going to proclaim some people good who aren't. But we want to minimize that kind of error.

And then what are the benefits of that? So, you know, if you're a parent I'd like to know if my kid gets one of these top 25 percent teachers or not. What's the impact on my child's attendance? We have -- this system gives you information on that as a function of how reliable the inputs into the system are.

And then finally, this report gives us a way of thinking about what is the quality of the evaluation system itself? How good are the decisions coming out of this? There's a very, very specific way of evaluating this numerically. So I think that pushing this agenda has the potential to focus discourse on the very questions that we have to focus on if we're going to transform this teaching and learning system that we have that has these really severe structural problems transforming it into something that's much more fair and productive.

MR. WHITEHURST: Thank you. We are open to your questions and comments. If you would raise your hand there will be -- someone will bring a microphone to you, I believe. Do we have microphone holders? Well, in a perfect system there would be someone to bring a microphone to you. But lacking a perfect system here at Brookings today I'll ask you just to -- I'll call on you and speak loudly. Go ahead.

MS. BARRETT: I'm Joan Barrett --

MR. WHITEHURST: Why don't you stand up, Joan?

MS. BARRETT: I'm Joan Barrett and I have this kind of Alice in Wonderland experience here. I'm thinking after this I am going to my physician and I'm glad that the legislators have nothing to do with how he's going to evaluate me. I mean, you said we chose this because states use it. Now, come on you guys. Why do you think states use it? It's because you guys have been pushing it all over the place. It's a kind of circle here.

I really am interested in this particularly in terms of what that common yardstick can be. And I was very taken by value-added the first time I saw it many years ago while many of you were still in graduate school, if not diapers, when Sanders first -- I first ran into him in Tennessee. So I'm not pooh-poohing it all. I think it is a wonderful thing to use as a single, to look at it and say what other evidence do I have. But to use it as the outcome to then compare all other evidence to, I have a lot of difficulty with this. So, I mean, I wonder would your system be different if you chose a different outcome and then looked at what predicted it.

MR. WHITEHURST: Well, our system is open to, and we're explicit about this in the paper, the system is open to other outcomes and is completely adaptable to those outcomes.

I think the question that has to be raised, we know that value-added and assessments are predictive of longer range outcomes, like high school graduation and college attendance and labor market returns. Again, far from perfect predictors but they are predictors. So I think the broader policy discussion about the other things that we measure is important and I think they have to be held to the same standard. You know, what do we learn from attendance? What do we learn from student -- surveys of students in terms of their engagement and aspirations? All of these things I think potentially can broaden our group or available measures quite considerably but all of them have to be addressed empirically.

I don't think, Joan, it's fair to say that this group has been pushing value-added or state assessments to the extent that these are widely deployed. They were deployed at about the same time that you say we were all in diapers and certainly preceded the value-added.

Next question, please?

MR. GOLDHABER: Can I say --

MR. WHITEHURST: Please.

MR. GOLDHABER: Joan, so I guess I just wanted to respond in two ways. Three ways. I agree with Russ that tests seem to matter for later in life outcomes and probably are quite highly correlated with some of the things in our own minds we think are important for schools to be doing. So the same kids that are doing well on test scores are also learning sort of the soft skills and are more likely to show up on time and be attendant to homework, et cetera, et cetera, et cetera.

But I also think that there's this issue of it's the only thing where we see variation. So there's this catch-22 because -- you're shaking your head -- but if there were other good data that were collected regularly in the same way that tests were, then there would be other things that we could look at to say, hey, you know, certain observations are associated with the way that principals rate their teachers. Or, hey, certain practices of teachers are related to the way that students receive the lesson. I mean, typically that kind of information is not collected so I think that there's a default. There's a default to value-added and that's an important context.

And the last thing I would say is that I would not think about the report. I wouldn't focus too much on the value-added aspect of the report, personally. I'm speaking for myself. I would focus on this report as a framework for how states and localities can think about the quality of their teacher evaluation systems no matter what your ultimate metric for success in a school system is.

MS. LOEB: Can I just add very briefly? I just want to second that last part because I'm actually a little uncomfortable with the yardstick of state tests on value-added but I don't think that's what this kind of approach to evaluating an evaluation system relies on. And you've got places like Florida right now that are trying to bring other things in. And I'm not sure what they're bringing in is any better than value-added but I think that we're in a time where the yardsticks are developing and hopefully improving. But it's nice to have a framework for how you would use them than to assess what we think of as very important in terms of it being a locally developed evaluation system.

SPEAKER: Thank you for the presentation. My name is Serfina.

I just wanted to pick on the lady's point. I feel uncomfortable as well with the value-added because I'm seeing the danger where we are going to have more tests and more tests. And I don't -- I've not come across anywhere where it says more testing leads to achievement. And that's going to be a danger when teachers have to compete to show that they are performing by the tests.

And I just want to think, looking at the systems that outperform us. For example, looking at Finland, looking at Singapore, Hong Kong, and the rest, I'm seeing a different thing that will make us think differently. They attract the best teachers into the teacher training and then they give them the best education. We don't do that. We just welcome anybody in and our teaching system of teacher education is so valued and some of them are not even up to quality. And I think that's where we need to be honest about how do we speak to the teachers' colleges to improve the course offerings that they do for our students.

And then when you look at the teacher evaluation, I'm looking at it more of a holistic point. You cannot come to evaluating if you didn't teach them how to teach it. So we need to start from the classroom, support these new teachers with mentoring and other things, and then evaluate them systematically or holistically so that we can improve. I'm seeing that this one may be -- I mean, it's my (inaudible) telling me that we're moving to more testing and more testing for our children which they need a break. Thank you.

MR. GOLDHABER: Okay, I can't help myself. I do think that people oversell the international comparisons and what we actually know from them. But I want to say something about the Finland, Singapore, et cetera. That is -- those are radically different systems in lots of different ways. So I might agree with you that wait, we ought to have a system that in principle teaching is very prestigious. It is a relatively stable job, high pay, and right from, you know, high school we start to funnel people into that profession. That model looks so radically different from the reality of education, the education training pipeline in the United States today that even if I agree with you I wouldn't even know where to start because it would be pure advocacy.

You know, we've got a system where teacher training and the requirements to become a teacher are delegated to 50 different states and in each state they're pretty loosely delegated to a bunch of different schools. So there are 2,000+ training institutions and just not the same kind of central control and not the same kind of second chances that exist throughout the education system at large in the United States. So if you're not a teacher, if you're not identified as a teacher in some of these countries, you know, by the time that you're 16, 17, 18, you really would have a very hard time ever becoming a teacher. So I think we both oversell the comparison and kind of ignore in some ways some of the benefits of having a much more open educational system in general in the United States.

MR. WHITEHURST: I would add we really don't -- empirically don't know much about what it takes to train somebody to be a good teacher. I mean, one of the things we are learning from this new administrative data and the ability to track how children do longitudinally in various classrooms is that the amount of variability and outcomes that are attributable to teachers is vast compared to our ability to predict which teachers are going to produce those outcomes at the upper end and the lower end. So it could be in terms of training teachers that we're in an intermediate phase where the ability to identify, you know, the top quartile or some portion of the top quartile reliably will allow us for the first time to figure out, well, what are those teachers doing that's different from what the rest of the teachers are doing and maybe then colleges of education could do a better job of passing on that skill set and knowledge set to teachers than they do now.

MR. STAIGER: Can I come back -- I want to bring this back to the framework that we have for evaluating and evaluation system. There's nothing here that says that an evaluation system can't put a lot of weight on, you know, the college you came from, you know, the program. You came out of LSU's teacher program. That one seems to be very effective. Or you came out of, you know, different tracks. There's nothing here that doesn't let that happen. The reason everybody, you know, everybody sitting up here doesn't jump to that is because so far we've never seen those things in the U.S. associated with teacher effectiveness very strongly. You know, Teach for America, which gets all these kids from Princeton and Harvard and all this stuff, they don't do any better than anybody else in the classroom. So, but that doesn't mean that it's not there.

So one of the beauties of this kind of approach is it starts to set up a system that says if you can start to flag people, say people who come out of a particular training program that's working, it's perfectly fine to use that in your evaluation to say this teacher is more effective. We know they're more effective because they went through this training program and that predicts their future performance. It's just right now that doesn't predict.

MR. WHITEHURST: Microphone?

SPEAKER: You mentioned some, thank you, federal bills at the beginning and one of them that you didn't mention is one that my boss has been helping out on the Hill called STELLAR. Have you heard of it? It's HR-1368. It was just introduced and it lays out requirements for states that do teacher and principal evaluations. So I recommend you guys look at it. It includes multiple measures, including graduation rates and observations. And it's interesting you talked about the attendance issue because that's required to be reported. Not required to be part of the evaluation but must be reported by teacher, by student. So that's part of it.

The other is that Colorado, one of the states you mentioned, passed a sweeping law last year on teacher evaluation and link to student outcomes and tenure. And one of the things they're struggling with right now in implementation is how much authority does a local school district have in setting its particular measures for teacher and principal evaluations. I wonder what your thoughts are as far as should it be statewide or should each district get to control those measures.

MR. WHITEHURST: In one of our previous reports on America's Teacher Corps, I think as Dan was suggesting earlier, I think part of the political science of that, if you will, or the political calculus was that, I mean, there are two ways we are going to get more teachers involved in meaningful teacher evaluation in this country. One is a heavy top-down approach. It could be mandated by the federal government or the state government as Colorado and other states are doing as suggested by earlier comments. And the other is more of a bottom-up approach. And our sense was, we're not opposed -- I'm personally not opposed to, in most cases, some form of top-down approach as long as it leaves flexibility. But we do think there are real advantages to having teachers be involved in the process of creating the evaluation system to which they are subject, particularly as that evaluation system is intended to help teachers become better teachers and not put them on the defensive and make them antagonistic to the evaluation system they're being subject to. So I believe consistent with our previous report we see a value in allowing teachers at a more local level to be part of the process of constructing the system to which they are subject.

MS. LOEB: Yeah, so I think for what we've done here with the America's Teacher Corps, we needed some kind of unifying way of evaluating evaluation systems so that they could be compared because this is supposed to come from the national. And so that's, you know, we spent a lot of time here thinking about what that yardstick would be and how important that yardstick is. But, in fact, a lot of our kind of underlying goal is the development of evaluation systems to in some ways professionalize the teaching force in a district so you recognize excellence and use it well and do all of those kinds of things. And for that I think you don't really need this yardstick. You really can have something as long as it's differentiated and can have a whole bunch of measures and be different across districts as long as, you know, the people there agree that it's what their goals are.

MR. WHITEHURST: Just one additional thought. One of the things that has struck me as I've been involved in this process with some of the smartest people in the country on the technical side of teacher evaluation is how little we really presently know about the best way to do this. And so, you know, I'm nervous about, you know, an imposition from above of a template for how this is going to be done, what percentage is going to be attributed to this and that part, until we allow some natural experimental at the district level and see what emerges from that.

You wanted to --

MR. GLAZERMAN: I just want to say, I mean, it sounds like all federal top-down things. It sounds like, oh, yeah, we'll say it and it will be done and everything will be uniform and it'll all be great. But so actually this top-down, think if you say we're going to have classroom observations and evaluations based on those. Well, everything we're learning from these rubrics, there are all these different ways. You know, I might use Charlotte Danielson, I might use Pianta, I might use -- you know, there are a million different ways to do this and I might implement it well or poorly in my district based on who I have there to do the evaluation.

So in some sense pretending that it's all uniform I think is a mistake, whereas -- and so part of our approach is that whatever you're doing you're going to get feedback on how well it's evaluating. And if you're doing it poorly you're going to get that feedback. No one is going to qualify for America's Teacher Corps or whatever it is and you're going to know that and you're going to be able to look back and say, you know, Cincinnati gets it. What are we doing different from Cincinnati or whoever it is? So we'll give that feedback on, you know, really the nuance of how you do these.

MR. RAUDENBUSH: I would also just point out that if you talk to district leaders in Denver or L.A. or D.C. or Chicago they'll tell you that it's not a trivial effort to develop and implement a teacher evaluation system and those are districts that are large enough that they'll have a staff, technically sophisticated, you know, and consultants and the ability to sort of do a lot of things that smaller districts can't do. So, just as a purely practical issue, there's going to be reasons why you're going to want policy at the state level to be directed at the state level just because of pure economies of scale. And that's why you see states like Tennessee becoming very directive and involved in teacher evaluation.

MR. WHITEHURST: Next to you. Go ahead. Yes.

MR. SIMON: My name is Mark Simon and in the interest of full disclosure I helped to develop the teacher evaluation system in Montgomery County, Maryland, that right now is in a major conflict with our state that has mandated that 50 percent of teacher evaluation has to be value-added scores statewide. Montgomery opted out. And so I'm interested in this phenomenon which I think Joan identified of the tail wagging the dog and, you know, I understand why economists and researchers love value-added. But from a teacher perspective and from a district perspective I'm wondering if when we look at districts and what constitutes a good teacher evaluation system if we see the same thing that you see.

So my question is as you embarked on this -- and by the way, I think, you know, Montgomery actually meets your standards but I think that in trying to meet your standards most districts will opt for a value-added kind of shortcut approach. So my question is as you embarked on this and you looked across the country, did each of you individually encounter districts that you thought had great teacher evaluation systems or much better teacher evaluation systems? And what were those districts? Because that to me is an interesting place to start rather than starting with schema that then can actually put good teacher evaluation systems in jeopardy when they're implemented. So, you know, what districts looked better than the rest as you looked around the country?

MR. WHITEHURST: I will say as I was responsible for writing a section of the report which required the location of examples of present teacher evaluation systems which allowed you to look at the association between the components of those evaluation systems and future student academic achievement. And it was extraordinarily difficult to find that information. So in the absence of that information, I don't know how to tell you which districts have great evaluation systems and which do not.

I mean, I think that's really, you know, a fundamental point in our report, that we currently lack a metric for comparing evaluation systems. And lacking that metric, how do you know? How do I know? On what basis can I accept your assertion that Montgomery County has a great evaluation system? It may but unless you can tell me that the scores that teachers get in that system predict something that you and I would agree is quite important in the future, then I don't have the confidence in your assertion and we need that kind of confidence and we need the ability to measure across districts. Maybe my colleagues have examples.

MR. GOLDHABER: I mean, I think there are a few districts that have been studied and people say -- I agree with Russ basically that we need some evidence. But I want to say something just about evaluation. You do evaluation either because you -- you do performance evaluation because you think the process of the evaluation, just going through the process, may make people better. You do evaluation because you think that you may be able to say something to individuals about how they can be made better. So it's not necessarily the process but the feedback that comes out of it. Or you do evaluation to figure out who ought to stay and who ought to go. And for the last two of those reasons you need variation in the evaluations. So not everybody can be excellent. And there -- you just see place after place after place that no matter what the evaluation system is, almost all teachers, 95, 98, 99 percent of them -- they just did a survey in Washington State -- 99.1 percent of teachers are at the top of their evaluation. Okay?

So unless you think that just being evaluated makes someone better, that seems like a very expensive process to go through to identify .9 percent of the population. So my answer to your question is that it strikes -- you know, I can't say that that's a bad evaluation system but I can say that to me it lacks a certain face validity because there's not a ton of evidence that just doing an evaluation makes someone better.

MR. STAIGER: Can I give a couple of concrete examples?

MR. WHITEHURST: Sure.

MR. STAIGER: Because it sounded like you wanted concrete examples. So one is not yet an evaluation system but is being developed through this Measuring Effective Teaching project, the MET project at Gates Foundation. And the kind of things that, you know, the preliminary report last spring -- so the first issue is it's always tied to how these things predict future value-added. So they have not yet broadened the yardstick but that's clearly part of the goal of the Gates project. But on that they're finding things like these student -- the preliminary report, the surprise was how well the student evaluations seemed to be associated with teacher effectiveness, future teacher effectiveness.

The practical one that I've seen in real life where there's been an evaluation is the Cincinnati one that's done by John Tyler and Eric Taylor and others, where they find a surprisingly strong -- they're doing I think it's a Charlotte Danielson modified kind of framework for classroom evaluation. Very intensive, person in the room, a lot of time spent there, and they find a very strong association with performance in future years on value-added again.

MR. GOLDHABER: But there's a caveat to that. Can I just say a caveat? The caveat is that they're looking in the subscores. And so it does look like in the subcores that what people are observing in the classroom, that it makes a difference. It predicts how well students are going to do in the future. But the same thing that you see elsewhere is true in Cincinnati in that most teachers when you look at the summative performance ratings are excellent.

MR. STAIGER: But you could use this information to predict.

MR. GOLDHABER: Sure.

MR. STAIGER: You’re right. The actual way it gets you is everybody is excellent.

MR. GOLDHABER: But I think that we need to think about it for prediction and for policy.

MS. LOEB: So just for -- I probably shouldn’t do this because I’m going to sound like I don’t support what we’re talking about here but here we go.

So I think the biggest concern for me about the value-added isn’t that the measures are that unstable or that I don’t think it should be student learning but that it really depends on what test you’re using. And so if you use a different test, teachers look different on it. That to me leads to big concerns about whether the state value-added is the yardstick we want to be using. I actually am really not sure that it is even, you know, with this face validity. But I do think that you want this differentiation and you want some kind of face validity. But I think the best we can do right now is face validity and we don’t really know what the right thing is. And so my guess is that your system is as good as another one if it differentiates teachers and that differentiation isn’t just by years of experience or something like that.

MR. WHITEHURST: We have time for one more question. Go ahead.

MR. STERN: Yes. I’m Barry Stern. I’m representing the Haberman Educational Foundation.

As if your task wasn’t great enough, what do you do when the nature of the teaching profession, in fact, is changing before our very eyes? Teachers are starting to work much more in teams. They’re starting to complement each other’s -- they’re actually hired as teams as opposed to being hired as individuals where they complement each other's strengths and weaknesses. They're using technology to improve productivity, increase diagnoses, and all that. So with the nature of the teacher profession changing where they are working as teams and using technology to improve productivity, how does that change the nature of your evaluation of teacher evaluation systems?

MR. WHITEHURST: The model we propose would be easily extensible to evaluating schools rather than just individual teachers or evaluating teams rather than individual teachers. It's quite generalizable in that way. I think the force of your question is to the extent that individual teachers cannot be linked to the performance of individual students it gets more and more difficult to evaluate them.

MR. GLAZERMAN: Yeah, I would agree. I think we encounter this all the time in talking with school systems and trying to go through this exercise of linking student performance to the teachers who are responsible for that learning, is we often find that there's a group effort. There might be a special ed teacher. There might be common lesson planning, co-teaching. All these complicated arrangements make it very difficult to capture adequately but they can be captured and in some cases you end up with a common performance measure that applies to all the members of the team and, you know, statistically we can't separate which individual teacher was responsible for what percentage of it but we might not need to. Or there might have to be some other mechanism for determining whether somebody is sort of, you know, the most valuable player on that team.

It's the same problem -- we actually have a set of papers planned for a presentation where it's going to be talking about value-added alongside a colleague who studies professional basketball players that have the same exact problem. You know, these people -- you know, how many points are scored while you’re on the floor but you don’t know if, you know, Scottie Pippen is a great basketball player because he’s always on the floor with Michael Jordan. And, you know, I know Chicago talk radio debates endlessly whether Scottie Pippen is a great player or whether he’s a free-rider on Michael Jordan. And you have some of the same problems arise in teaching. But as Russ was saying, these are very surmountable and they just require us to make the teacher performance measurement system flexible and adaptable to the way instruction is delivered.

MR. RAUDENBUSH: Let me say something slightly different. I think this question raises an important point about what the unit of evaluation should be and I think Russ made a good point which is that one might orient oneself to evaluating schools. And the rationale for that would be to give everybody in the school an incentive to help each other become more effective. So there is an argument that if you want to promote team play of this type you should reward the entire school and that internal to the school you should, you know, there would be an incentive for people to create ways of helping each other improve their expertise. And in fact, in some cases where someone is not improving their expertise, the team would have, you know, would want that person -- in fact, I can give you an example of where what happens in a school that operates exactly this way that you don’t have to fire people because they leave because they know and everybody else knows that they’re not responding to attempts to make things better.

So I think I would prefer to allow enough flexibility. In other words, rather than saying we are committed to individual teacher evaluation, I would be committed to information systems that can be used to dramatically improve the efficiency and fairness of the school system. And I think that some combination of school and teacher evaluation is going to get played out and it's going to depend on how you organize your reform efforts. I mean, that may seem too wishy-washy but I think the exact same system could be used for schools. I mean, you could say let's identify the top 25 percent of the schools and you could ask exactly the same question about how good is the evaluation system. Maybe we can only identify three percent because our evaluation system is bad. I would personally be very much open to that as an alternative. I'm not sure about my colleagues.

MR. WHITEHURST: I want to thank you very much for your attendance today and your great comments. We learned a lot from what you had to say and ask about. Thank you.

* * * * *

CERTIFICATE OF NOTARY PUBLIC

I, Carleton J. Anderson, III do hereby certify that the forgoing electronic file when originally transmitted was reduced to text at my direction; that said transcript is a true record of the proceedings therein referenced; that I am neither counsel for, related to, nor employed by any of the parties to the action in which these proceedings were taken; and, furthermore, that I am neither a relative or employee of any attorney or counsel employed by the parties hereto, nor financially or otherwise interested in the outcome of this action.

/s/Carleton J. Anderson, III

Notary Public in and for the Commonwealth of Virginia

Commission No. 351998

Expires: November 30, 2012