Q&A with Douglas Harris
November 1, 2013

Q: Could you give a concrete example of how the screening approach would look and work at a school, both in terms of the measures and how they are integrated?

A: I would like to be able give an example of a school district that is doing it, that would be the most concrete thing. But here is how I would imagine it working if I was at a school and I had to evaluate my teachers. I would look at all of the information from prior years. You would have to have some sort of data system for looking at the prior information, including the most recent value-added measure, and either create a weighted average or use the matrix approach, some way of determining which teachers are below a certain threshold. You might say if a teacher is low on either their current value-added or their prior classroom evaluation, then you would observe that teacher four times during the school year. For teachers who are high on both of the metrics, you might only observe them a smaller number of times, maybe once or twice. That is where the cost saving comes from, that you do not have to spend as much time on the teachers you have greater confidence are being effective. The key choices are how much additional information you are collecting on the teachers that are in these extreme categories and what information are you using in the first stage to decide who you are putting in those categories.

Q: The MET report discussed different weighting schemes for test score gains, classroom observation, and student surveys. They advocated 33-50 percent weight on achievement gain. Where did they get these numbers and how would you response to the MET report in this regard?

A: What they are doing is trying to find out the best predictor of future value-added. They are taking value-added as being the goal. Looking at current information, they looked for the best combination of current measures to predict future value-added. So that is where their weighting schemes are coming from. That is an important caveat, that they are still focusing on valued-added as being the thing they are aiming for. Ideally if we are doing these kinds of studies, we would have some true, perfect measure of performance and we would be validating against that. Instead they are validating against future value-added as the metric. Given the constraint that no one has a perfect, true measure of performance, I think what they did is a reasonable thing. But it does not necessarily mean that those are the optimal weights if you want to be thinking about teacher performance more broadly than just contributing to student achievement.

Q: What would evidence of accuracy of these approaches look like? What research would be needed to support the accuracy of these approaches?
A: The first thing that we would want to do is to apply all three approaches to a single set of teachers and see whether the answers even end up being different. The hard part is determining what true performance looks likes and we do not have that. But if all three measures resulted in the same categorization of teachers, then that would be a pretty good indication that accuracy is not really the main criterion we need to worry about. If they all lead to the same answer, then they all must be equally accurate. That would be a first cut. I would want to see whether the methods yielded the same answer. If they did not yield the same answer, it would be harder. You would be back to the problem as I mentioned with the MET study, that we do not have that perfect measure to work with. From there the analysis would get a little bit more difficult.

Q: Is that research that could be done with existing data?

A: Yes. I think you could do some simulations on how it would work. The piece of the screening process that is not available is the idea of using value-added to improve the observations. That part you would miss with existing data. If you set that part of it aside, you could do it with existing data. For example, say that you have four observations for teachers, just pretend that you only had two and randomly select two of them and play out what the performance categories would look like under those different hypothetical scenarios.

Q: The use of student learning objects (SLOs) have been increasing. How might value-added results be combined with SLOs? Can you offer any specific examples?

A: That's a great question. I have a discussion of that in the brief, even though I did not mention it here. I think that the same logic applies to SLOs as applies to classroom observation. Both of those are focused on the quality instruction teachers give to their own students. So, you could do similar analyses. If you wanted to improve the quality of the SLOs, you might use that correlational approach. Although I do not know of evidence on what those correlations would look like, there may be some evidence out there, but I don't know. The problem there is that sometimes you are using SLOs when you do not have test scores, so it could be hard to do the comparison between the two if you are only using SLOs in the non-tested grades and subjects. That said, I would think that the same principles would apply here as would generally apply with classroom observation.

Q: Under the screening process, if a teacher was found to be average or proficient on a screening measure, let's say value-added, would you recommend no evaluation or observation for the current year to lower costs and time? Why or why not?

A: I would recommend that every teacher is observed at least once a year to avoid the feeling of a scarlet letter. Teachers are going to know who is being evaluated and who is not. I think that having some teachers who are not evaluated at all seems to be instinctively too extreme. I think you also want to be setting an ethos and precedent that everyone can get better, no matter how good they are. I think that having at least some evaluation for every teacher every year helps to establish that.

Q: You mentioned that it is important to assess the statistical properties of the measures. In your brief, you say that there is a step before that to establish the value weight. How would a district go about defining what they value?

A: So it is not the easiest thing to do, but there are three main steps. Step one is thinking about the objectives of the school. Broadly defined, what do they want their students to know and be able to do when they leave the school? It is the basic philosophical question about what the school is there for. From there, then you have to think about how you might measure school and teacher contribution towards those outcomes. The first stage may be deciding how much of it is academic achievement versus citizenship and social skills. The second phase would be to identify the measures of academic achievement. You could use value-added there. Classroom observation may be more focused on things beyond academic achievement that the teachers are doing. You would need a different measure to capture that other goal of what you want students to be able to do. Then you have a mapping of what measures you need. That leads you to what the value weights should be for measurements since you

would be making judgments about how each of those measures captures the goal that you have established for the students. The final step is thinking about the validity and reliability of the measures. If you think that they are all equally valid and reliable for measuring student progress towards those goals, then you can probably stop there and just worry about the differences in the scales. If you have one measure that has a scale that is zero to 100 and another that is zero to ten, you do not just want to take an average of them because that will not capture the true contribution of each of them. I think that captures all of the key considerations.

Q: A lot of districts are currently just measuring a small set of measures of teacher quality, but if we were to add additional outcome measures that we care about, how would that impact our thinking about the accuracy and relative merit of these approaches?

A: I think that if additional measures are capturing different goals of education, then it would affect which approach you use. To that point I made at the end, if they are measuring the same things, the screening approach makes relatively more sense than the others. If the additional measures you are bringing in measure different goals entirely, then it would make less sense to use the screening approach. If you had two goals and two measures of each goal then one approach would be to use screening. You could use one of the measures for one goal and one of the measures for the other goal as the first stage in the screening process. And then use the other two in the second stage of the screening process. It is not that you could not use the screening method, it just becomes a bit more complicated in certain situations. There is nothing that definitely determines which approach you should use.

Q: On the issue of overlap, you mentioned that when measures overlap, there is less reason to collect multiple measures. You make this point that it is important to understand whether measures capture different or the same elements of effectiveness. How might we begin to figure out whether these measures are capturing the same or different constructs of teacher quality?

A: One thing that you can do is some basic statistics. How correlated are they? If they end up being very highly correlated then that will give you some indication that they are capturing the same thing. What is a very high correlation it is not any easy question in itself because these things have a lot of measurement error, so correlation between value-added and anything else is almost never more than about 0.5. But that is the first thing you can do. Another part of it is logic. In classroom observation you have four or five specific criteria, like for example in the Danielson Framework. Those parts are labeled. For example, one part is supposed to measure classroom management for example. In some cases, especially with classroom observation, it should be reasonably clear what it is capturing. It does depend on how well those categories align with how you are seeing the goals you have for students. That part can be tricky. Since classroom management is not related to academic achievement any more than it is citizenship and social goals. You would want classroom management for all of those things.

Q: A follow up on the screening question. Besides the medical screening model, are there other out of industry examples of the screening approach for personnel decisions or examples of the matrix approach?

A: That is a good question. I am guessing that there are, but I have not looked for them, so I do not have any examples off the top of my head.

Q: What do you think about more complex statistical analyses, like principal component analysis, for this purpose? Why are these types of analyses not being used yet with multiple measures?

A: I don't know for sure they are not being used. I think there are some analyses that have done things like that, though I cannot remember specific studies off the top of my head. I think that the approach does make sense. It is a way of identifying the independent information that each measure is providing. To the degree that the concern is about multiple measures capturing the same thing, the principal component analysis does make sense. I was talking in somewhat simpler terms, but the more sophisticated analyses would probably be better.

Q: What about teacher accountability towards society, teacher towards parents, teacher towards students, apart from contribution towards student test scores?

A: The other approach, and the MET study did some nice analysis of this, would be the student surveys, (the Tripod surveys) where the students are asked about what is going on in the classroom. Those turned out to be reasonably good predictors of value-added, even more than classroom observation. I think those can be valuable. Surveys of parents can be a lot harder. Trying to get a response rate that is reasonably useful is much more difficult with parents. It is easier with students because they are in the school.

Q: You mentioned that what we are really concerned about is the validity and reliability of personnel decisions, but we only seem to discuss decision making models based on algorithms and quantitative measures. There are other decision making models out there that emphasize expert intuition, for example naturalistic decision making. How do others decision making models compare to the status quo?

A: Good questions, I am not an expert in decision science. There is certainly a debate, even a public one, within education about how much discretion there should be in the evaluation of teachers. We are moving toward is a more mechanistic approach. You are put in one of these categories and a decision is made based on what category you are placed in. It is taking the discretion away. I know that there is a lot of concern about doing that. The reason it evolved this way is because of the way decisions were made in the past and the fact that almost every teacher was in the highest category and received very little feedback on how they were doing. So it seemed that when judgment was the determining factor and the system was not so mechanistic that the system did not seem to work very well. There seems to be fairly broad support for the fact that it did not work. That is not to say that this new approach is that much better, but there is a gut reaction to giving too much judgment for fear that it would go back to the way it was before.

Q: You mentioned that a different kind of evidence is required to understand how teachers respond to accountability measures. What kind of evidence are you referring to and how might states or districts collect it?

A: For me, what we are trying to do here is improve teaching and learning. It is not easy to measure those things. I think that is what that is the larger conversation about, where value-added is just one piece of that conversation. There is reasonable amount of consensus about what good teaching looks like. I would not say complete consensus, since different people have different goals. There could be

disagreement on the degree to which you focus on academic content. Though there is some measurement error, I think that most people with expertise in teaching, even across the political spectrum, would walk into a classroom and have a fair amount of consistency on what a good teacher looks like. I think that we need good measures of what is happening in the classroom and look at how the accountability is changing what is happening in the classroom. That is the different kind of evidence I am speaking of. We don't have very good evidence of that. I think that we will be able to get it now that we are doing more regular classroom observation. To the degree that evidence and data can be used to do additional analysis to be able see how quality of instruction is changing overtime. That is really want we want to know. What is really happening in the classroom that is different? Is it getting better? Is it looking more and more like what we want it to look like? We are getting a lot closer, but we are not there yet because it requires some sort of before and after comparison. You have some outcome information before the accountability system and you have the same information after. Do things seem to be getting better over time when these new systems are put in place?

Q: We have a question on tradeoffs between the matrix and weighted average approach with regard to the resulting personnel decision. Placing a teacher with good observation and poor value-added or vice-versa in mid-category may not give the right rank, but may give right policy outcome, that is no effect based on available information. The result would be that the teacher is not sanctioned or maybe not rewarded.

A: It is possible that that is true. I think that it depends on whether we think that each of those measures is accurate. My sense, and I have seen analysis like this – where you have a valued-added measure and a high value-added teacher who gets a really low classroom observation or vice-versa – is that one of the measures is just wrong. Trying to minimize the degree to which any measure is wrong is important. If that is the case, then averaging them together is not very useful. Keeping them separate is a flag for people to think about which might be wrong is important. If they are measuring completely different things then they may both be accurate. If one is measuring contributions to the school community and the other is measuring contributions of teachers to students, then those could be more reasonably and consistently different from one another. But to have the classroom observation be consistently different from the value-added, then probably one of them is just not right. That is where the concern comes from with the weighted average.

Q: How exactly does this integrate with Race to the Top and for those states that have mandated measures? If you could speak to districts and states that are going full steam ahead with their current teacher evaluation systems, what kind of advice would you offer them?

A: I think that this is the most frustrating part about writing a brief like this. I know that probably half of people who read this thought, "Well, I like that idea, but I cannot do it, so too bad." I took a long-term perspective in this. The federal government has not codified this and said that we have to do this. The states have. The states and the districts that have gotten Race to the Top have written into law how this will look. But we at least want to get some more experimentation going, where it is possible. Maybe in the states where more flexibility was built in or where exceptions can be made or something else can be tried to see if it works differently. I am not sure the screening process would yield different results or be seen as more credible by educators, which is also important. But the only way to find out is if someone tries it. I think there is a plausible enough case that these other approaches might be better that we

should be trying to experiment. The state laws make that harder, but if you think that one of these might be better for your situation, if you can lobby for it, maybe you can get there over time.