Q&A with Douglas N. Harris

October 31, 2012

Q: You talked about the issue of multiple measures, but given that the measures themselves have a relativity weak correlation with each other, what does that suggest about the kinds of patterns of evidence that districts are likely to see when evaluating a teacher?

A: This is something districts are grappling with. Districts get one answer when using one measure of teacher performance, but because the measures are weakly correlated, they get different answers with different measures. When you look at multiple measures, what do you do when you get conflicting info about teacher performance from different approaches? The first thing you wouldn't want to do is panic and say one measure is clearly wrong, it's possible they're both right. And it's also very likely that there is a reliability issue here, and that the next year the measures will look more similar for that teacher because there is some random error this year that makes the measures look more different than they really are. But I think in those cases, and what I have suggested elsewhere, is to think of multiple measures as a part of a process rather than a single number you are trying to boil down to determine performance. If you do see conflicting evidence and you are making a high stakes decision, try to collect additional information. If you see a teacher performing poorly in a classroom observation, but in value added they look good, I would suggest doing more classroom observations to make sure there is a conflict and to make sure you have the information you need before making a decision.

Q: When you speak of reliability and validity with respect to value added, do you mean the statistical models are not reliable or valid, or the assessment to test student achievement? Can you tease that out?

A: Value added is a kind of a “garbage in, garbage out” situation. Whatever test scores you put in the value added model affect what comes out. In other words, value added can be no better than the student achievement tests, both those things are part of validity. What we’re interested in is drawing inferences about teacher performance, and there are multiple ways that can go wrong. It's possible it goes wrong because the scores themselves are wrong and are not capturing what we think is important. In addition, the statistical model may not be accounting for important differences of students in different classrooms that also affect student performance, and that's a selection bias problem. For validity purposes, both of those pieces are relevant.

Q: Are you finding more research directed to other measures of teacher effectiveness now that so much attention is being paid to value added?

A: Yes, the Gates MET study has a lot more ongoing work that they have not yet released. They have the data to assess, for example, the reliability of the classroom metrics. On the validity side, we have to think of the test validity of the other measures in different ways than what we have done with value added. I think you could say something about the validity of the principals’ observation in the classroom based on the process itself. For example, a valid process could be one in which the raters have to go through a standardized process to learn to assess correctly, are practicing that on videos of instruction and are making sure everyone is evaluating in the same way. If you didn't have a process like that process, you would certainly be more worried, and you would likely get different results just because of rater effects. I think there is certainly interest in it, and interest in developing other measures in a similar way to how things evolved with value added. In the beginning we weren't as focused on issues of validity with value added, and only now we are trying to think of creative ways to test validity. Now we are in the same spot with these other measures. There has been a rush to create the measures and get them off the ground and then test validity after. Maybe that's not the right order of things, but that seems to be the way the world works. So I think there will be a lot more attention paid to the issue of the validity of the other measures as we go. One general concern I have in this area is that there is a misunderstanding of why the measures differ, and one of my intended contributions of the brief is to help people better understand why they might be different and interpret differences between measures and their validity.

Q: What do you think about the validity of SLOs and the relationship between SLOs and VAMs?

A: We do not know as much about SLOs or how to set up the process in a way to help ensure validity. There is a larger question about how to think about validity of SLOs given they are inherently not standardized. So it becomes a lot harder to think about validity when we intended it to be different for different classrooms and different teachers. It will be a lot harder to establish validity criteria and think about it differently with SLOs than we think about it with other measures that were intended to be more standardized.

Q: What are your thoughts specifically concerning special education teachers being evaluated with value added models?

A: One thing we’ve learned, particularly in the past few years, is that value added works better in some circumstances than others. I'm working on a paper that looks at how value added works in middle school versus elementary school. Most evidence we have on the validity of value added is from elementary school, but it looks like there are lots of good reasons to think it doesn't work as well in middle and high school, at least in the way we are structuring middle and high school tracks. Tracking in middle and high school is quite different from elementary schools, and in the way that the single standardized test aligns in different ways with the classroom instruction of different tracks. Special education is another example where you'd have to be pretty worried about, for example, the test aligning with what students are learning. If they're not at grade level, what they're learning will not be captured in the test and if it's not well captured in the test, the value added will not work and you will get a misleading picture of the effectiveness of special education teachers. So that suggests moving to something more around classroom observation around direct measures of practice rather than getting it from student outcomes. Another approach is to look at an alternative assessment and to redo the value added with alternative assessment - special education students still have to take those in the vast majority of cases. I think we know a lot less about whether that's going to work. We have a wide variety of student needs and the special education category is a very broad category, we certainly don't know much about it and there are reasons to be worried that even the alternatives assessments in those scenarios aren't going to give you a valid assessment of teacher performance.

Q: Can you say more of what is meant by value added, and how does that relate to the quality of teaching?

A: I didn't go into what value added is earlier for the sake of time, so I think I might refer to other publications like my book and others that lay out what it is. The basic logic is to try to identify what teachers contribute to student test scores and the big challenge of that is that different teachers have different students, so trying to account for what the students bring to the classroom is a challenging task. The intention is to identify the contribution to student academics as measured by the test, and the question we grapple with today is how well they accomplish that and how well they accomplish that relative to other measures. With value added we define performance in a particular way that is very focused on academic skills, and very focused on how those skills are captured in the test.

Q: How do other types of growth measures such as simple growth, student growth percentile, etc. relate to teacher observations and other teacher evaluation components?

A: There are two ways to look at that question, one is how do value added measures as typically calculated compare to those other ways of looking at student growth. The answer seems to include two things. One, how you account for prior student achievement is by far the biggest adjustment, for example in my book I compare the level of test scores at the end of the year and compare that to the simple growth measures where you subtract the prior year's scores and look at the average scores, you can get very different results when you make that very simple adjustment. And then what value added is doing is going beyond that simple change score to adjusting for a bunch a different factors, testing for student demographic perhaps or whether they're a special education student, those addional adjustments can make some difference but not nearly a difference that the initial adjustment of prior student achievement. Then there is student growth percentiles, the work I've seen on that so far suggests that student growth percentile gave a pretty similar answer to a simple value added measure with no demographics or other covariates, one of the distinguishing features of student growth percentiles is that they don't account for prior demographics, although they recently came up with a way to do that but generally the ones in the field do not account for these prior student demographic, so you end up getting a fairly similar answer to a value added measure. So given how similar the different growth measures are, that also implies that all those system measures will be similarly correlated with classroom observation. I haven't seen studies that have done that but if the correlation with a growth measure is high that is also going to imply that the correlation between those various growth measures and non-growth measures will also be about the same across the growth measures.

Q: Do you see a need to differentiate definitions of teaching effectiveness in disciplines when students perceive their performance in these assessments important to their ability to compete for jobs, admission to college, internships, etc?

A: There are a lot of factors that affect student scores other than the teacher. Student motivation is a tough one because it is reasonable to think that part of the teacher's job is to increase motivation. Motivation is partly what being a teacher is about. But part of motivation isn't determined by the teacher and it is determined by external factors especially the home environment, community, etc. The evidence supports the idea that tests reflect student skill in different ways for different students, there are racial components to that, and motivation level going into the exam is part of that story. And there is not much we can do about it except in some degree that issues like motivation end up being captured by prior scores. If a student is consistently not very motivated to take a test that will reduce the level of their scores but not necessarily change the growth of the scores – that consistent lack of motivation will be reflected in the prior score and that will be accounted for in the value-added measure. I don't think we know how well it does that, and this is a value judgment; I don't know how much of student motivation we want to put in the responsibility of teacher when we define performance.

Q: What do you think of the effects of the Common Core implementation on these observed correlations you discussed earlier, what do we think we will see with this new infrastructure rolling out?

A: Part of the answer is, each state is going to be different in terms of how well their prior tests and standards align with new standards and tests, so it is conceivable that within a state that it doesn't change all that much because everybody to some degree has a distinct disadvantage. Everybody in the state is switching to new standards. They are not necessarily all switching at the same time, I know some districts have tried to get a head of the ball and make the switch over before others have and that could distort things. I think we also don't know what affect it will have when we switch tests. This is something that happens in many states, many times. Every 6-7 yrs many states will decide to revamp tests, although maybe not as radically what the Common Core will. These changes will make year to year comparability more tenuous. This is certainly a reason for concern. I think a lot of districts are struggling with switching to Common Core and layering on teacher evaluation problems on top of that. This is quite a hurdle and one that everyone is struggling with. I think we can expect to see the correlation drop at least in the short term while were making that transition because you can end up with more random error, but in the long term I wouldn't expect too much change in those correlations. It also depends on the level of random error, part of the random error comes from the tests themselves. It is really unclear if the new tests will be more reliable than the old tests and if they are more reliable then that would generate higher correlations with classroom observations rather than lower correlations.

Q: What do you think the actual impact of Common Core assessments will be on understanding the performance of small groups? Will this help us look at the differential performance of minority groups?

A: I don't think it really helps from a value added standpoint, because even though we are going across states, the problem for identification of value added to specific subgroups isn't that we don't have enough students to compare, the problem is that each teacher has too few students in those subgroups. From a value added standpoint, even if we had one test for the whole country it wouldn't necessarily solve that problem because none of the teachers would many of those students so you'd have the same problem we have now, which is too few of students in the subgroups to say much. You'd have information for program evaluation, but that still wouldn't solve the problem of teacher-level evaluation.

Q: Is there a possibility that these value added measures could be different by academic subject?

A: There is some indication that value added measures work better in middle school and math, relative to others. I say this based primarily on the study I did with Tim Sass where we looked at returns to teacher experience. There's pretty strong reasons from literature on worker productivity in general that people get better as they get more experience and that pattern when we look by subject and grade tends to be clearest in middle school math. We see that gradual but diminishing increase in teacher performance more clearly in middle school math more than other places. That partly counters what I said earlier the fact that middle and high school are different and in some ways problematic because of this tracking issue, in that respect it's harder to say. I will clarify so everybody realizes that when we compare teachers in reading, we're only comparing within reading. There is no way to say if a teacher teaching reading is better than a teacher teaching math. There is no way to make that comparison. In math, all we're doing with these measures is saying where a teacher stacks up in the distribution of performance on math, and likewise for reading. Another reason why I think math might end up being an easier subject to use value-added for is that we have fewer examples of other teachers and other external factors likely to influence the scores. One early concern for value added was that some students read over the summer and some students read at home and that was going to affect their reading scores and that was something teachers couldn't control. Also social studies classes in high school involve a fair amount of reading that could influence reading scores, and you don't have much of that in math. There is less math going on in other subjects and less math going on at home than reading overall. You can say fairly confidently that value added would work better in math than other subjects. By grade level, it is a little harder to tell if there are advantages and disadvantages to elementary versus middle school and high school to tell to make an overall assessment about whether it works better or not.

Q: Do you think value added has been influenced by Race to the Top and are you concerned?

A: Value added was the basis for Race to the Top in a lot of ways, so we would not be having this conversation about measurement of teaching if not for value added measures pushing the envelope. If I interpret the question to mean the influence of high stakes on the validity of value added measures, there is certainly reason to be worried about that. Teaching to the test and possible cheating can result from high stakes and that will reduce the validity of any measure, including value added. The pace of implementation for Race to the Top has been very fast and you've got a lot of new value added measures entering the space.

Q: We have real cases where a student scores high in math and struggle in reading, how important is it to include reading scores when you're looking at performance in math?

A: We have done work in that area and it doesn't seem to make much difference. You would usually include prior math score in that situation and the question is what happens when you also include prior reading scores. The reason why I think it doesn't seem to make much difference is because, while there may be cases in which students are high on one and low on the other, there is still reasonably strong correlation between reading and math scores on the average. There may be some random error in the equation, but the strong correlation between reading and math scores is why I think it doesn't makes much difference in terms of the math value added measures to add reading scores into the model.