Q&A with Doug Harris
June 7, 2013

Q: You talked in detail about the differences between elementary and secondary in terms of the reliability of value-added. We also know that many states and districts are using these measures for other school-based professionals, including nonteaching staff. What can we say about the validity of evaluating other school-based professionals, for example, speech and language pathologists? They have a role in the system, but don't have the tight linkage in this student-teacher class system.

A: This one comes up a lot. There are a lot of people in schools who aren't the typical classroom teacher. The general principle, for me, is that we should be holding people accountable for what they can control. For speech and language pathologists, for example, there's not an obvious test score that goes along with what they're doing. That's the first problem. The second problem is that those instructors aren't spending much of their day with those students. It's likely that the outcomes of the students that they're serving are going to be driven by other teachers. So, I would suspect a value-added measure using a standard state test for a speech language pathologist is not going to be very accurate. What to do about it is a harder question. The intuitive response for a lot of places is to come up with a different test that's specific to what the instructional leader does and that's how we'll solve this problem. That would potentially address some of the problem. But, it's not clear that we have good tests for that, or what that's really going to capture. If we don't have good tests for that then we shouldn't do it. How do we evaluate them? There are other ways. We're already in these new teacher performance systems using classroom observation, and other additional approaches – using a combination of information to judge their performance. I think that in this case, if you don't want to do the additional test route, which I suspect would be a questionable one, simply relying on some sort of observation of their practice would probably be a better approach. Another thing you could do is have team-based value-added. So, rather than trying to assign the student achievement just to that teacher, and possibly have a group of teachers that are primarily responsible for all of the language skills of the students, a value-added measure would be developed for that whole group. The disadvantage is, if you're trying to individually evaluate one of the teachers, that's not going to work very well because it's really measuring the performance of the group. The information might still be useful for general improvement purposes for the school. So, there are different possibilities, but none of them are great options. We need to get away from the idea that we need to have individual value-added for all. There are a lot of reasons of why that would be difficult to accomplish and probably not very helpful.

Q: One of the questions that comes up a lot is has to do with teaching at the edge of the distribution. We have a question about teachers of high achieving kids. They're often concerned that value-added doesn't capture their contribution because the tests don't challenge their students. I think there's a question about the test ceiling in general, but is it even more of a concern in a system where there's pretty substantial tracking?

A: Yes, I think it is more of an issue with tracking. At the elementary level, to some degree, you have some ceiling effect problems because some schools just have higher performing students. So, you might be worried that in those schools value-added won't work very well. At the secondary level, now you'll know for sure that at every school you'll have some students that will be higher performing in those upper-track courses. Again, that effect is being counteracted by the fact that those advanced track students seem to give the teacher other advantages. So, the disadvantages, in terms of the ceiling

effect, are counteracted by the fact that those students seem to have a higher expected growth curve anyway. We need to be paying attention that there are two different problems offsetting each other. The ceiling effect is part of it and is no doubt an issue, but it's being counteracted by this other effect at the same time.

Q: We've got a question about interaction effects as well. Have you taken a look at the interaction effect between content and grade level on the predicted value-added results?

A: The problem is that we don't observe content very well. To answer that question you'd want to go into the classroom and see what's taught. I don't know if anyone has done it, but that would be the way of trying to get at this issue. I end up talking generically about content and that there's probably this mismatch, but it's hard for me to say anything specific about it because I don't actually observe the content.

Q: There's a question here about differences across subjects. Did you see substantial differences in reliability between, say math and ELA that we've seen at the elementary level?

A: That's a good question. I don't remember off the time of my head what the difference was between subject. I gave the 25-50% range in a slide, and part of that range is based on the fact that it wasn't the same across grades and subjects. In general, with these data, reliability was noticeably better in math versus reading.

Q: Would you have the same sort of concerns in a state where a growth model, rather than a value-added model, is being used?

A: I don't think it would be different. In both situations, you'd still have the same basic issue of sorting and alignment going on. We're assuming in all of this that we're talking about ignoring the track. So, is ignoring the track more of a problem in a growth model versus a value-added model? It seems to me that the problems would be pretty similar in both. Another reason that I'm saying that is that those two measures tend to be reasonably highly correlated. Value-added measures and growth measures tend to be fairly well-correlated. I think it's going to be a problem in either case. The question is how well you can account for the tracks explicitly in the model.

Q: The Measures of Effective Teaching project by the Gates Foundation gave us a fair amount of data about the correlation between value-added and observation measures. Is there anybody that's done that at the secondary level that you know of? Are there any studies at all that would give us a correlation between some sort of a growth measure and teacher observation data?

A: There are a few studies. I did one with Tim Sass where we looked at which observation was getting the principals' assessment of the teacher. I don't think we noticed results that were too different across grade levels. The correlations are somewhat higher in math and this is probably partly due to the higher reliability of math value-added noted earlier, which reduces the measured correlation with any other measure. The difference in these correlations was not as great as you might expect given the large differences in reliability, so there are probably other reasons for this as well, such as principals' ability to math versus reading instruction.

Q: We have a question about the practicality of your advice about assigning weights based on reliability. Do you know of any districts or states that are putting that into practice, where they're using reliability to think more critically about the weighting strategy?

A: Implicitly, I think we are doing this to a certain extent when we use shrinkage. I don't know how familiar that term is going to be, but it's become a pretty common practice. When we adjust for shrinkage, at least some methods for adjusting for shrinkage would implicitly do this. But it wouldn't just be shrinkage based on the number of students. It has to be the form of shrinkage that's accounting for the whole standard error, not just simplistic adjustments based on the number of students. It would have to be accounting implicitly for test measurement error and those kinds of other factors that affect the standard error.

There are several questions districts and states could ask. One is, which shrinkage technique is being used? The second is whether it's the kind of shrinkage measure that's actually going to account for this problem. I think most shrinkage measures would account for this. So, even though it's not explicitly weighting based on reliability, implicitly that probably is going on in the places that are doing shrinkage.