Carnegie Knowledge Network • What We Know Series on Value-Added Methods and Applications

Webinar 5: How Should Educators Interpret Value-Added Scores?

Q&A with Stephen Raudenbush

April 5, 2013

Q: Are there other ways to think about increasing the precision of value-added estimates? In addition to just increasing class size, what else could we do to increase precision?

A: We are starting to see people use more frequent, and typically diagnostic, assessments. This is one way that you can augment the information. Let's say kids are being assessed throughout the year. We're watching to see whether the kids are moving. We're aggregating more information over each child and aggregating more information across for the whole teacher, getting a bigger, better picture of what's going on. I think the repeated assessment using these diagnostic assessments could increase precision. Then you also want to combine that with multiple, highly trained observers coming in and looking at what's going on. There are some other more novel and, maybe not as tested but still interesting ideas about collecting student work and tests, and looking at the instructional quality. So, there are a bunch of things we can do, but the things I'm mentioning require that there be a school that's functioning well, doing these assessments and observations. It's not clear to me that this can all be done at the level of the district. Those are the things that I've thought of. I think the Measurement of Effective Teaching study has shown that you can get more information by augmenting test scores and value-added with observations of these students.

Q: To what extent do you think you could base value-added calculations on using multiple assessments within the same year to actually yield a higher correlation? So, you aren't comparing children under different circumstances with different teachers.

A: Great point, because you're collecting data in same year you're still talking about the same collection of kids. That still could be unstable from year to year, but at least you're getting a more precise estimate of what's happening during that year. You'd actually be getting a learning rate within the year. That's one thing you don't get with value-added. In the situation you raise, you'd get a picture of what's happening with the kids that's much more detailed. You get a better picture of that year but, if next year's kids are going to create a different chemistry all together, there's still going to be some instability. That wouldn't be solved by just having multiple assessments throughout the year. For that, you'd have to aggregate data across years to try to try to see, what the persistent component is of teacher effectiveness.

Q: This actually leads right into the next question. So, what about the number of years? There are various arguments in literature and in the public sphere about how many years of value-added one actually need to make judgment on teachers. Do more years yield a better understanding of an individual's productivity?

A: I believe you'll get more reliable estimates with more years of data. Even though there's instability across years, you're aggregating a lot more data. The test of truth will be to construct exactly the same graphs I was just showing you, we call them caterpillar plots. You can construct a caterpillar plot for an average value-added taken over two or three years. What you'd expect to see are those error bars to contract and those confidence intervals to get shorter, thus being able to more clearly distinguish between teachers at different parts of the distribution. What I would say, in answer to your question, is that I want to see it. I'm an empirical person. I want to see the caterpillar plot you get when you calculate value-added after two or three years. I'm not sure we've seen many of those yet. Maybe some of the other experts you've been talking to can show us those caterpillar plots, but I haven't seen them. You would expect them to get more clear as you gather more data. The main point I'm trying to tell everyone, make sure you see the caterpillar plot. Don't let anyone con you into looking at a single number.

Q: We've heard several warnings from you Steve. The issue you raise about the danger of making bad decisions given the existing system for measuring productivity, years of service and that sort of thing. What do we actually know about the precision and reliability of the other measures? What do we know about the reliability of these sorts of instruments over time?

A: I'm thinking about Ron Ferguson's Tripod survey, when you ask kids questions about their teacher, and when you aggregate over all the questions that you ask all the kids in the classroom, you get reliability up in the neighborhood of .8 and higher. So, that seems to be the single measurement device that I've seen that has the highest reliability. Now, you might say, classroom observation is going to be even better. The problem is there are two sources of error in classroom observations that you have to worry about. One is that different raters are using different standards, even when they've been very carefully trained, you saw this in MET, the Measurement of Effective Teaching project at the Gates Foundation. They've done this in thousands of classrooms, and very meticulously. You need to have at least four different raters look at a classroom. You need to go on at least four different days. Now, I'm not saying sixteen days overall, I'm saying one person goes in on day one, a different person goes in on day two, a third person goes in on day three, and another one on day four. I've done analysis even before MET, so we came up with exactly the same conclusion. But, still, we're going to get reliability in the neighborhood of about .60-.65. We're not going to be up around what Ron Ferguson gets us with the Tripod. So, because raters have different standards and, even when you train them, they see the world a little differently. Also, the day you go in and the time of day can have a random effect. In other words, what you see will differ if you go in on Friday or Monday, or if you go in the morning or afternoon. These things bounce around. We call it temporal instability. To make the classroom observation system work, you need to go in more than four times and you should have at least four raters. Then you still get a modest level of reliability. It's very useful information but, even with classroom observation, you still want to see the caterpillar plot, and you're still going to have uncertainty. Now, when you start combining classroom observations, students surveys, and value-added, then you start seeing more precision.

Q: That's an interesting point. So, the Measures of Effective Teaching study laid out how to think about combing these measures. We've got a question here about the use of value-added estimates to better inform observation results or vise versa. What sort of guidance would you offer for how to actually think about doing this?

A: I think it's going to matter a lot what kind of decision you want to make. If you even think about the caterpillar plot we're looking at here. If I want to learn something about the teachers having the most difficulty, the very troubled ones, I have pretty good information on this. At least I know they aren't up at the top. The same thing with the people at the top. We can see them being distinguished. I think that when you start putting the information together from multiple sources, and you see, let's say if you're looking for people at the bottom and you see people are showing up at the bottom on two or three measures, you're starting to get some very strong evidence about those people. It's a lot harder to distinguish those in the middle. That's one of the lessons of the caterpillar plot. And even if I start combining information, and if I looked at a caterpillar plot where I was looking at the average over all the information that we've got, it would still be somewhat difficult to make any kind of strong statement about people in the broad middle. We could make stronger inferences about those people at the very bottom and the very top.

Q: How effective is observation by itself at identifying the bottom 25% of teachers? If you think about order of operation, is value-added better at more effectively identifying those?

A: I would say they're roughly in the same ballpark. When reliabilities are in the neighborhood of about 0.5-0.6, you're going to see caterpillar plots look like the ones I showed you. So, when you look at the reliabilities of observations in MET, they're pretty close to the same neighborhood of value-added reliabilities. As I mentioned, if you look at student surveys, you get a significantly higher reliability. Then there's another problem. I'm very skeptical about using students to make high stakes decisions about teachers. I think for professional development and formative purposes that's great, but as soon as students are brought into the high stakes, we might corrupt that indicator. The idea of getting information from students is a reasonable idea, particularly, and I think it's in Ron Ferguson's work, as soon as they get into 3rd or 4th grade they're giving some pretty good information.

Q: Are the studies of the use of student surveys for teacher evaluation based on their use in high stakes settings?

A: Actually, they're starting to be used for stakes in Chicago, where I am. I have been advocating against it. I worry that if kids know that what they say on the questionnaire is going to have implications for the future of the teacher, it's putting them in a funny position that we don't want to put them in. But, people are doing it. Maybe I'll be proven wrong. I don't know whether the results are in on that though.

Q: Do you know of any research that actually uses the Ferguson approach in a high stakes setting?

A: We're trying something and have no idea whether it works or not. We know the student surveys are very good when you use them in a low stakes setting, but whether they're going to continue to work well when we put them in a high stakes setting is another question. We just don't know the answer, but that's the way education is. People just do things on a mass scale without knowing about whether it's going to work.

Q: I've got a technical question here. What actual statistic are you referring to when you're talking about reliabilities?

A: It's more or less a Cronbach's Alpha. The way I define the ratio of true score variation to total variation is essentially the same idea. It's a little more complicated because what we need here is a multi-level model. What we've got is variability between items within children. How many items we have tells us how reliably we can discriminate between the children. But that's not really our goal. Our goal is to aggregate the children, where we're regarding the children as informants, if you will, or raters of their classroom. We're aggregating over them, so then we have to take into account the variability between kids, and then we have the variability between teachers. Typically, we also remove the school variability, or not -- that's another question. In this case it doesn't matter much whether you remove the school variability. So, we used a three level model in order to get all of the variance components separated out. It's kind of a fancy form of Cronbach's alpha, based on a three level model with items within kids within teachers, and then four levels if you look within the schools.

Q: How do you recommend districts communicate the imprecision of value-added? In many cases districts are making decisions that include value-added as one component of a high-stakes teacher evaluation system. How do you message the value of value-added without undermining the credibility?

A: If you're talking about a collection of teachers then you need to see the caterpillar plot. I just don't see any way around it. And if you're talking about groups, you need to report the anticipated error rates—the anticipated false identification rate. That's what you do in medicine. They have the same thing. If you think about a medicine, you want to know the false positive and false negative rates on that. Take the test for prostate cancer. It's got a false positive rate. A lot of people who claim to possibly have cancer do not. It's been a huge controversy. Doctors and patients couldn't have a good discussion unless they were recognizing the imprecision of measurement. We need to have the same discussions in education. Why shouldn't we? It might be embarrassing to the school district to tell people the truth about the degree of imprecision in their measurement devices, but we need to have that as a norm. Even if we have imprecision, it doesn't mean that the information is useless. It does mean that we have to take it into account and might have to get some additional information.

Q: One of the other things we're seeing at the state and district levels is the combination of multiple measures into composite scores. What impact does variable reliability of the input measures have on that? What's the impact of adding more and more potentially noisy measures to a composite?

A: It depends on how correlated they are. If the inter-correlations are positive and of a reasonable level, then aggregating more information is going to give you more precision. If the correlations are relatively weak, somewhere near zero, we might not be getting anything out of adding the extra information. We have tools to do that and we can look at the caterpillar plot for any aggregation as well as for any component thereof. Generally, these things are reasonably correlated. I mean, generally, we do get more information by compositing these things. Another issue that's come up in the literature, if you have multiple measures, it's a little harder to game the system. I'm thinking about Atlanta, sorry for bringing that up. If you have one indicator, and it's an all-powerful measure with high stakes attached to it, people are going to game it. But, if you have multiple measures, and you're not putting all of your eggs in one basket, it's less likely. I'm not talking about flat out cheating. I'm talking about teaching to the test or learning how to get good results on one specific instrument. Multiple perspectives are more likely to get at something that's more generally true. I'd say this is even true with testing. Using different tests, different assessments that have different formats, is probably a good idea because we don't want to have a bunch of teachers that are just teaching everybody, until they're blue in the face, how to teach multiple choice tests.

Q: This question is talking about the Common Core. Most districts will be switching to new assessments. I guess we're assuming year-to-year correlations will likely to fall during that period when states and districts are shifting from prior state tests to some new tests, potentially using different constructs, more sophisticated measures. These would be correlations between the old and new tests. Based on your concerns right now, should we be worried?

A: When a new testing regime comes into play, it should be interesting to see whether the rankings of the teachers are perturbed by that new test. If it's highly perturbed, it could be because everybody is so acclimated to the old one, that is, they're gaming the old and teaching to that particular format. It could also conceivably be that the tests are testing different content. It's a tough thing when you change because it's harder to compare across, from the past to the future. Presumably, you don't want to do this too often, but we'll see how that goes.