Q&A with Daniel McCaffrey
December 14, 2012

Q: Can we use value-added when student test scores are not vertically aligned? If two scores aren't vertically aligned, aren't you comparing apples to oranges?

A: One of the most common methods used to calculate value added is to take historic information on students, predict how they will do given their historic background, and then compare their actual scores to those predictions. The reason why this method is so popular is because it doesn't require tests to be vertically aligned. Vertical alignment might make the predictions stronger, but as long as you can control for the differences you'd expect to see across classes, you don't need the tests to be vertically aligned.

In many ways, value added does not directly measure "growth." These measures use prior test scores to predict how a student will do on this year's test, based on how they did in the past. In a sense, that is a form of growth, but not necessarily in a pure growth metric, so it may not be exactly measuring growth. Now, even Student Growth Percentiles (SGPs) that have growth in the name don't require vertically aligned tests. SGPs find students who have performed similarly in the past and find where a student ranked relative to every other student who performed similarly in the past. That's the student's percentile. Using a single test, if we had a group of 4th graders who scored 100 points in 4th grade and then looked at how those students scored in 5th grade, the student who did the best in that distribution is at the 100th percentile and the student at the bottom is at the 1st percentile. Again, even that which has growth in its name is really growth relative to expectation instead of a pure growth measure. Some methods try to use a pure growth measure with vertically aligned tests, and they sometimes rescale the tests across time. Even those methods measure growth relative to the distribution.

Q: Are the issues of confounding the same between Student Growth Percentiles and value added, or is there some benefit to one strategy over another?

A: SGPs and value added use a very similar notion at the heuristic level, because of this, SGPs suffer from some of the same problems as value added. If you look at SGPs that are fit with only the prior year score and examine their relationship with the percent free/reduced lunch at the school or the average achievement of students in the class, those relationships can be very strong and not that different, if we look at the simple value-added models I was talking about. There is a paper on the SAS website with some evidence on this, as well as several other papers coming out that look at this.

The other distinct difference is that value added sometimes tries to control for other factors such as free/reduced lunch or race. I know that in some cases it is controversial whether or not theythese factors should be controlled for. You can sometimes see differences between the methods that control for demographics and ones that don't. In particular, some value added models control for classroom aggregates, like percent free/reduced lunch, and we have seen that they make big differences on how strongly value added will be related to the average achievement in schools or the average achievement in the classroom.

Value added gives more flexibility and more ways to control, the percentile method has other advantages, in a sense it has some flexibility advantages and percentiles are something people understand, but it can suffer from the same potential for confounding as value added.

Q: What are the best types of models to use for high school end of course exams?

A: Generally, the same value added model is used for high schools that is used for elementary and middle schools. But the devil is in the details in high schools, even more so than for middle and elementary schools. For the method to work really well, you need to capture the differences among how you'd expect students in different classrooms to do. Let's create a cartoon. In one school I have two groups of 9th graders, one group taking AP calculus and the other group taking algebra. I'm going to say their 8th grade math tests will account for how well the students will do on a 9th grade test that covers material from algebra through calculus. Clearly, we don't think the 8th grade math test will do a very good job of predicting the huge differences in those students and all the other ways they are different. We can run a value added model that controls for their 8th grade score, even if it isn't vertically aligned, to try to control for those differences, but we have to really understand that that model is lacking and not rich enough to tell us all the ways those students are going to differ. We need to understand what else made the students in those different classes different. When we get to high school, we face the issue of very big differences among classes.

The last time that we have universal tests is usually 7th or 8th grade. These are generic tests that are not specific to certain courses. When you move into high school, students take more specialized courses (different types of science courses, more specialized math, and even variation in English courses), so you can fit the models but they won't necessarily have the achievement. If you are trying to base predictions on weak predictors, you get the case where confounding creeps in. You get back toward the simple model problem. If you don't correct for enough things, you leave a lot behind and what is left is systematically different across classes. You wind up with the opportunity for confounding; attributing differences in student characteristics to teachers. Especially in high schools, we need to fit the value added and see if it looks like confounding, such as students in advanced classes doing better and students in remedial classes doing worse.

Q: Could PSAT exams serve as a good pre-assessment for high school value-added models, or are they likely going to suffer from the same general-specific problem you mentioned?

A: The PSAT is a test that is, in some places, broadly used. It is given later than 8th grade tests, and its content is different from an 8th grade test. It may be more strongly connected to what a student is doing. I don't know because I have not seen data on how well the PSAT really explains the difference you will see among the students in their outcomes. Among students in the same classroom, I don't know how much the PSAT explains the variance in outcomes; I don't know how well it will function as a predictor. I think it is an empirical question, but it is an example of the type of question that districts can evaluate themselves to make a decision.

Q: What do we know about the other measures that are being deployed, e.g., SLOs and observation metrics, and how they might help us with confounding?

A: I don't think there is enough research yet to say. We have a dearth of information in general in high schools, and of the other kinds of measures, we just don't have enough rigorous research on how they function in general. We particularly don't know much about their use in high schools. For example, we don't know how sensitive something like an observation measure is to the differences in classes. We have a study on this issue using algebra, where we have 8th graders in middle school and 9th graders in

high school taking algebra. We see that the 8th graders are higher achieving using an algebra pre-test score than the 9th graders, and we find that the 8th grade classroom observation measures tend to be higher than the 9th grade classes as well. This could be because they are the really elite students who get the better teachers. It may be more of a focus in the middle school, or it may be possible that the middle school teachers who teacher algebra really work harder at it in a way that is different from the high school teachers. We don't know the exact reason. But this is an example of a case that has the kind of confounding we'd be worrying about showing up in the data. We really can't sort out differences between student-level characteristics and teacher contributions.

Currently, we don't have any evidence on the other measures that leads us to believe they sidestep the issue of potential confounding. Because we have standardized tests and standardized methods of calculating value added, we have been able to scrutinize and rigorously evaluate value-added methodology, and we put it through the ringer in many ways. These other kinds of measures are newer, and measures like SLOs are often much less uniform, so we don't have the ability to analyze them with the same rigor as we have with value added. Some of the same challenges are likely to be there, if we're comparing across different classrooms. If SLOs can be restricted to more similar kinds of students, like SLOs for AP calculus, that might help avoid some of those problems. It may help to have the same denominator.

Q: Is the value-added technique being used successfully with local assessments?

A: There are some places using value added on truly locally created exams. Hillsborough County in Florida and a county in Maryland are developing a lot of their own assessments, scaling them, and then trying to use them for value added. One thing with locally developed tests is that those data don't tend to be available for people to analyze as rigorously as they've analyzed the data on the state tests. There are some cases where people have looked at comparisons where they have taken value added from one state test and compared that to value added estimated on another state test. For example, in addition to the state test, Houston also gives the SAT10. A group of researchers estimated the value added on those two tests in Houston and found that value added is not that strongly correlated across those different tests. The MET study also found that value added is not the same from one test to another. It is not clear what that means. We might expect the tests to be capturing different aspects of teaching and learning, so we might expect the scores to be different. One other thing about value added is that we want to be careful how broadly we want to generalize results from one particular test to inferences on a more broad set of achievement measures.

Q: What kind of guidance can you give to district and state agencies who are thinking about appropriate uses of value-added for teachers who don't have tests associated with their grades or subjects?

A: I don't have good insight into that because I focus on places that have tested grades and subjects. I know some places are looking at making attributions from the whole school to teachers without value added. The only thing in that space that comes to mind is an experiment done in New York where they gave schools bonuses based on how well the school did. This is not exactly the same, but they did look at rewarding people based on the school-wide performance, and whether that affected student outcomes. It was a random experiment and they found that it didn't. I don't know if it tells us that assigning other teachers different teacher's value added is wrong, or if it tells us that giving bonuses based on student performance doesn't work. I don't think there is a lot of evidence out there that says

that this is the right way to do things. Like I said, most researchers are studying the properties of value added only in places where we have it.

Q: Back to your point about the composition of schools and the degree to which schools are homogenous in the type of students they serve. You used the phrase "avoid comparing teachers in very different situations." What would that look like if we implemented that advice in a district?

A: It is a really good question, and a fair one. I don't have as good an answer as I wish I had. One thing to do would be to group similar schools together. I'm not sure exactly how to create the groupings, or how fine or coarse the grouping can be, but a lot of districts, for various reasons, classify schools based on the population they teach. You could restrict value-added comparisons to teachers teaching in similar schools. There are challenges with that. You could have a condition where you only compare teachers in lower income schools. If, in general, all the teachers in those schools are not doing well, then you give teachers a free pass by comparing them to a mediocre group. You might want to do other things to get a sense of whether those schools are producing well in other ways. So when you compare teachers head to head, you may want to start with comparisons of teachers in similar schools and then also check if, as a group, they are performing well using some other methods, or you may need to use other measures on teaching. The other idea that is sometimes suggested is comparing teaching within a school, and that may be useful for certain purposes, in particular for principals who are thinking about how to encourage professional development and how to assign teachers in a school. But there are definitely problems with that, such as increasing competition within schools and limiting collaboration. That is something that needs serious consideration, but it really has to do with how you plan to use the measures. We don't have strong evidence to date whether it creates competition or limits collaboration, but there is reason to worry about that.