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

Webinar 8: Will Teacher Value-Added Scores Change when Accountability Tests Change?

Q&A with Daniel McCaffrey

August 14, 2013

Q: Since a teacher's value-added score is relative to other teachers in the study unit (district or state), and all students in the unit will be exposed to new tests, would we expect value-added ranking to change that much?

A: They can. When tests change, suppose the distribution moves up or down. In New York, the distribution moved down, but not all students move down all the same amount. If all students don't move the same amount, then classes wouldn't all move down the same amount. In every one of these empirical examples I talked about, they used value-added where they were all relative to other teachers. But, when you ranked teachers on one test and ranked them on another, they all shifted around quite a bit. If you think about how the tests align with what's happening in one particular class, how those particular students do, or what that teacher does, they wouldn't have to move the same way. Maybe if I pick an extreme example it would be easiest to understand. Suppose there was a teacher who figured out how to really teach exactly what was on the test very well. They had a test that was really focused on certain content, and every year they had very similar items on the test. This teacher really practiced teaching those items, so those students were really good on those items. When they change the test, that teacher's students aren't going to do very well on the test because she doesn't really know how to teach the materials, she knows how to teach the test. That teacher would change in the ranking quite dramatically. That's really a cartoon to make the point that the rankings don't have to move in the same way. On a lesser scale, that could happen in many other ways and that's exactly what we're seeing in the examples that I presented earlier.

Q: We've got a question here about the test technology. Is there any research that would shed light on what's going to happen when we move away from what is mostly fill-in the bubble technologies to adaptive tests? We've already discussed change in content, question type, and difficulty, but we're also introducing the technology of adaptive testing.

A: I'm not aware of specific studies that have really gone in and compared estimates on adaptive testing and non-adaptive testing. There are some adaptive tests currently being used. The one thing that we do understand, from a theoretical standpoint without that specific empirical test, is that adaptive testing does improve the reliability of the test scores for students. It improves pretty dramatically for students on the low or high achieving ends of the distribution. The non-adaptive tests are much less reliable for those students than for students that are near the center of the distribution. So, by improving reliability of student tests, that should help to improve teacher-level value added. Any way that we can improve the quality of the student-level data, we're going to improve the quality of the value-added. It might help reduce some of the error in the value-added. It might help reduce some other kinds of issues and features of value-added, in terms of how well we really can control for students' prior achievement, things like that. In these ways it will improve the value-added. I don't know if it will result in big changes in teacher rankings.

Q: Do you think we might also introduce some instability through students being exposed to different content provided through the adaptive algorithm?

A: I don't know. Even though the items are different, my understanding is that the items meant to cover pretty much the same content but presented in ways at the right difficulty level for different students. Going back to the previous question, if there are ceiling or floor effects on tests we might see bigger changes in value-added if we switch to adaptive testing. We do see floor and ceiling effects in certain states where we don't seem to see much variation among students at the top or the bottom of the distribution. You kind of get to a certain point where the kids just all get the lowest score, or you get to a point where they're all kind of piling up at the top. One nice thing about the adaptive testing is that it really does remove those kinds of things. Switching to adaptive testing we might see teachers at the top or the bottom with more dramatic changes in the value-added. This is very much along the lines of the anecdote that the one district administrator told me about.

Q: Some researchers have suggested only using teacher value-added when you have a minimum of ten students to achieve a minimum level of reliability. Do you know where this threshold comes from and could you make a recommendation for a threshold that states should use?

A: I know where the idea of a threshold comes from. I don't know where the ten comes from. You'll remember in the slides I was talking about the notion that there's a lot of error in tests. That error makes value-added change from year to year. That error can contribute to value-added on one test being different from another test. A small number of students just really doesn't tell us much about a teacher. We all understand this point very intuitively. If you look at one teacher who has one student we really wouldn't think we know as much about that teacher as we would a teacher who has 100 students. So, people came up with a minimum of ten students. I think that's a business rule that states are using. I hear 10 come up often, and I think it's a real compromise more than a specific statistical rule. More is better than fewer, so we want a minimum number of students. In some research, the instabilities really accelerate once you get below 10. You do see some leveling out at ten, although 15 is better than ten, and 20 is better than 15. But I think the reason people pick ten is because it's a compromise point. Below ten is really unstable. We have a lot of teachers who only have ten students who are tested with prior scores and the other requirements. If we required more than ten students, we might start to lose too many teachers. The desire to have quality value-added for as many teachers as possible led to this choice of ten. I don't have a better number than ten. If someone wants to pick a threshold, I think working through these kinds of decisions rules on balancing the desire for good estimates from large samples versus not losing too many teachers is how to pick that threshold number.

Q: Is there anything out there in the research literature that talks about what happens to value-added when the test stays the same but we have changes in curriculum and standards? Some states are actually experiencing that right now. They've rolled out Common Core standards and new curricula but are still taking their traditional state assessment. They haven't moved to a new assessment. Is there anything we can tell that can guide us about what we should expect?

A: That's a really good question. I'm not familiar with any research that specifically looked at that. A lot of value-added research I'm familiar with is often used with data that's somewhat old. The time sensitive information that directly connects to what the standards were at the time of testing isn't always there with the test score data. I'm also not sure if there are as many examples as dramatic as changing from the current standards to the Common Core standards. I know are some places where there are some really big changes going on. In other places maybe those changes are not as dramatic. I don't have any information on that and I'm not sure we have any real historical incidents that might line up particularly well with it.

Q: Will Common Core assessments be using criterion-referenced scores rather than norm-referenced? What are the implications for using criterion versus norm-referenced scores for doing value-added?

A: That's a good topic. I'm not exactly sure. Usually when you talk about criterion-referenced tests, you don't think of them being on a developmental scale because they really are specific to a criterion at a particular level, at a particular grade, and so it doesn't really make sense to think about putting them on a scale. I know people have been working away from doing that and I know that there is some discussion, but I don't know where they've landed at this point, of the new consortium tests having a developmental scale or a scale that goes across grades. Not sure whether they're going to focus on criterion- or norm-referenced tests. I think where the key issue for value-added is not norm or criteria referenced but how well tests from a prior grade serve as control for the differences in outcomes you would expect among classes, just because of the differences in the students assigned to them as opposed to differences in teachers. The whole idea of value-added is to use prior achievement to say, once I control for prior achievement, and some other things about the students, then I can act like Mr. Smith's class is the same as Mrs. Jones' class. So, that any remaining differences are really about the teacher. The big difference is that with criterion-referenced tests, usually the material is specific to a grade and the controls from the other grades are often not on the same scale. You can't directly use them in a way that you would assume they were on the same scale. Norm-referenced tests don't have to be that way, they're often more likely to be put on a common scale across grades. Other than that kind of difference, I don't think there are really big differences in the way people use tests for value-added. There are things done to get around the scaling issues, like using student growth percentiles and the median growth percentile approaches. With those approaches people try not to rely on scaling assumptions quite as strongly. Those methods tend to get fairly similar results as the more traditional value-added methods.

Q: You gave us one example of the effect of test change on a subgroup with the example from the impact on high achieving students, pre- and post- test change. Do you know about the impact of a test change on other sub groups?

A: No, I can't think of any off the top of my head. There's a lot of work looking at how value-added may work with sub groups. I know people are very interested in students with disabilities and English language learners. I think that work is looking at how those students do, how those tests do. I can't think of any particular examples where they would look at a test and say, did it change what was happening with students with disabilities. It is a really good issue. I think with the Common Core tests, it's really particularly important because, with computer adaptive tests they're going to be trying to bring in more students. They're not going to have the alternative tests for students with disabilities. I know that changing the accommodations that are given to students can have really big impacts on students. You would think that if they started to have really significant impacts on students, at some point they'd have to create instability in the value-added. It would be hard to believe that some of these things won't show up with differences for, say, special education teachers. One other thing, lots of teachers with special education students have very small samples anyway, so their estimates are usually more unstable to begin with.

Q: All of that is imbedded in your recommendations of taking a hard look at subgroup performance or at the ends of the distributions, right? That's the recommendation you gave us at the end of your slides.

A: Exactly right. Those are just the kinds of things that people would want to do – to really make sure that the change in tests didn't results in really low or may be high value-added for teachers who have lots of students with disabilities. If they find that the change does disrupt ranking, they might want to come up with a reasoned response that might offset the disruption and the potential for people to conclude that the value-added estimates are problematic.

Q: This question goes back to your discussion about your informal poll of value-added providers where you discuss controlling for prior math and reading, that is, including the off-subject prior assessment score as a control. Will controlling for both math and reading always strengthen the value-added estimate? Is your recommendation to always use off-subject tests for value-added?

A: The answer is that using more tests is always going to be beneficial. It may be of very little benefit, but it is. You can cook up an example where you have highly specialized high school tests and now you're going to throw in the kid's third grade reading test when you're trying to sort of control for student Algebra 2 test scores. Well, that probably won't do much to help you. But, in the context of, should we always use reading and math, or should we throw in social studies? I think there are advantages to using additional tests score. There's sometimes a price to be paid if those additional test scores are not completed by all the students. If it starts to introduce trouble in the estimation because more students don't have test scores, then it might lead you to not want to include them. I think with the analyst I was talking about, he uses math and reading from the prior year. That way students who have reading in the prior year almost always have a math score. So, you don't introduce much data loss by adding more tests from the same year. I know that in Bill Sanders work at SAS, he would always want to have three prior tests, three prior years in the same subject or, for third graders, he would try to have three prior third grade tests. I'm not sure three is exactly the right number to have, it's not a bad number. The notion is that when you don't have much information to control for how students differ across classes, having more tests and pulling together additional information will help. Test scores across different subjects are fairly highly correlated, so that always gives us some additional control by bringing them in. Where you might decide not to use more tests is if you're going to lose a big sample because lots of kids don't have a certain kind of test.

Q: Is it always the case that you should use one more additional measure? How should people think about it? What's the guidance for including additional measures?

A: Suppose I'm a firm believer that the content on the state test is what I'm interested in. I believe the state test is a very good measure of that content. I really want to understand teachers' contribution to student outcomes on that state test. What we found in MET was, pretty much what you need for that is a couple years of value-added, you don't gain much from pulling in other information. What we learned from other studies was, if you want to improve what you learn about that, probably pulling value-added across multiple years is your best bet. The other information from observations and survey improves things a little, but it really doesn't add much. Where multiple measures seem to be more important is in a case in which I believe the state test is important, but I also believe that there's different content and there's more to what a teacher does than just contribute to growth on that state test. In that context, bringing in multiple measures has more benefit. Our research gave us some indication that weighting things more equally made them more beneficial for predicting teachers performance on students' growth on tests other than the state test. That makes sense. If you want your target to be kind of broad, then pulling more information together will give you a better chance of being close to multiple different things. So, should you always combine? It's not clear you would always need to combine. We didn't find any times where combining was detrimental, but we did find times where combing wasn't particularly helpful. Most measures of teaching have positive correlation among each other. A teacher who has very high value-added is generally not typically the worst teachers on another measure. So, there is some information from almost any of those others measures. So kind of like adjusting for prior tests, including that extra information does give me a little more to know about the teacher on any specific aspect I want.

Q: I guess the one thing that would figure in is that some of these measures are quite costly. If there's only a moderate gain, the costs may outweigh the marginal benefit.

A: That's right. I was assuming that data was already in hand, but if you have to collect the data, then there may have to be a decision whether the additional measure provides enough additional information to warrant the cost, or the disruptiveness. Sometimes we're thinking actual cost, in terms of salary, but something like surveying students can be disruptive and take a lot of time. There are also other kinds of negatives. You can picture there may be a negative response to certain kinds of measures that you'd want to avoid. I can't think of a specific example of such a measure. Things like observation and classroom measures seem to be included, at least in some cases, because they seem more desirable. Going back to the previous question, I was assuming the data to be combined all met some minimal level of quality. I just remembered an example where this might not be true. I was working with one state which was using with SLOs (student learning objectives) as one of the measures of teachers. The way the SLOs worked in that state is that many teachers did not meet their student learning objectives. Teachers made the objectives too hard. The information they had to set the objectives was so poor and people just didn't have the experience with setting standards to develop the measures that were really feasible. I'm not sure that those SLOs gave accurate information. They could potentially harm people because it's not really clear they represent what a teacher is doing with her students as much as how lucky she got with writing a reasonable SLO. They are also unstandardized so comparing across teachers may create errors that are not similar. I think if had a choice to use or not use those data at this point in time, I wouldn't use them, given just how little we know about what they measure. So, to revise my earlier comment. If you have of measures that are so poorly defined and built that they could have negative value maybe they should not be combined with other measures.