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

Webinar 4: How Stable Are Value-added Estimates across Years, Subjects and Student Groups?

Q&A with Susanna Loeb

February 13, 2013

Q: You mentioned that some VA models are less stable than others – can you talk about the characteristics of ones that are most stable and why?

A: Generally, value-added models that contain a lot of fixed effects tend to be less stable because they narrow the comparison group of teachers and, in so doing, reduce the variation used to produce the value-added estimates. This reduced variation can lead to less reliable estimates. When you include school fixed effects in a model, you are essentially comparing teachers to other teachers within the same school. When you include student fixed effects in a model, you are comparing teachers who have served the same students. Models that do not include school or student fixed effects don't narrow the teacher comparison group; instead, they allow for general comparisons to all teachers. In making these general comparisons, we can control for the characteristics of the teachers' classrooms and their schools in place of the fixed effects.

With that said, we don't know which model produces value-added estimates that are more accurate measures of teacher effectiveness. The general conclusion of the research literature so far is that the reliability, or the amount of error in the estimated with student fixed effects may not be worth the benefits of the narrower comparison group for comparing teachers in more similar situations because they introduce substantially more error.

Q: I worry a lot about data quality in my role at a state agency. As many as 30% of our teachers' rosters are not fully accurate. Will solving this problem help us with the stability issue?

A: The validity of value-added estimates as measures of student learning in particular teachers' classrooms clearly relies on the accurate assignment of students to teachers. If students are attributed to teachers incorrectly, teachers' value-added estimates will not reflect their students' learning or their teaching effectiveness. Some roster problems result not in students assigned to the wrong teachers but to students not assigned to any teachers. This discrepancy also can create problems. Value-added estimates are more accurate when they are based on more students. If teachers have few students assigned to them for the purpose of creating value-added estimates, these estimates will not be precise.

Q: Do you have any guidance about the appropriate correlation for different policy uses? For example, a 0.5 correlation might be adequate for description, but is that enough of a correlation for awarding bonuses and making employment decisions? Are there any guidelines about a bar you want to be across before you make higher stakes decisions?

A: Essentially, the answer is no. There is no bar. It depends on what other options are available. If there are other sources of information that are better, more accurate and reliable, for the goals that you have (e.g. predicting future effectiveness) then you don't want to use value-added measures, but if those sources don't exist, then you do. Most likely, value-added will be best used in combination with other sources of information, such as observations and subjective assessments from various constituents.

Q: In your transition matrices, the top and bottom quintiles seem to be the most stable while the middle moved around a lot. Why is this, and what are the implications for policy making?

A: Yes, that's right and not surprising given the errors in the measurement and the positive relationship between value-added in one dimension and value-added in the other dimension – for example, value added in one year and value-added in the next. A teacher in the top quintile in one dimension can only be in the same or lower quintile in the second dimension. Even if this teacher is better in the second dimension, she or he still cannot be in a higher quintile. Similarly, a teacher in the bottom quintile in one dimension can only be in the same or higher quintile in the second dimension. However, a teacher in the middle can move both up and down.

There is some information in the middle. For example, the teachers in the third quintile, on average, will be significantly more effective at raising test performance than teachers in the second quintile. However, while the averages differ, if we took one teacher from the second quintile and one teacher from the third quintile, their value-added scores probably would not be statistically different. This inability to distinguish individual teachers but the ability to distinguish groups of teachers means that policies that use value-added measures to group teachers will lead to the misclassification of teachers. The benefit of the policy or practice will depend on whether the benefit of using the information on the average differences outweighs the costs to individual teachers of the misclassification.

Q: How much will truancy and attendance issues at the student level affect a teacher's value-added score? Should attendance be taken into account?

A: If students rarely come to class, teachers have little ability to affect their learning. As a result, it may be inaccurate to include students with low attendance in value-added measures or, alternatively, we may want to adjust value-added measures for student attendance. In practice, it is possible to adjust value-added models for attendance in the same way these models adjust for prior test scores, students' family income, or other characteristics of students or classrooms. Some value-added measures used in research do adjust value-added for student attendance, though they tend to adjust for student attendance in the prior years, not in the year that the student was in the teacher’s class. The reason for this lagged adjustment is that teachers can affect student attendance. If a student in not engaged in class, he or she may be less likely to attend. Engaging students is part of teaching and, as such, the value-added measures should capture this. By controlling for prior attendance, value-added measures can adjust for students’ tendency to attend school, not their actual attendance which could be influenced by the teacher. The drawback is that if a student has particular issues in a given year that lead to reduced attendance but are out of the control of the teacher, then lagged adjustments will not do enough. It is a tradeoff.

Q: On sample size, how small is too small when it comes to the number of students assigned to a teacher for calculating his/her value-added?

A: It is a decision that districts, schools, and states will have to make. The studies that I've looked at tend to use about 15 students per teacher as a cutoff, but this is an arbitrary cutoff. Sample size matters both for the precision of teacher value-added estimates and for the number of teachers (and the number of students) included. Sample size considerations, and measurement error considerations more generally, will depend on how the measures are used. For example, if the value-added measures are used solely to determine where kids are learning and where kids are not learning, then sample size matters less; I would use it for any number of students. If I only had data for five students in a given class, I'd still like to know how much those students are learning. However, if the measures are going to be used in a formula evaluation with meaningful consequences for teachers, then I would be cautious about small sample sizes. Value-added measures are imprecise and, in general, the smaller the sample the more imprecise. This imprecision calls for caution in their use and for combining them with other measures.

Q: It was clear from your presentation that two years of data are better than one. But I think what we heard you say is that there are diminishing returns after that. Is there a notion of best practice for how many years of data to include when calculating value added?

A: The benefits of the third year of data are not as big as the benefit of a second year of additional data but they do add some predictive power as do additional years after that. Adding a year of data increases the ability to predict future value-added but decreases the number of teachers for whom this prediction can be done. You'll have to balance the precision benefits of additional years of information with the costs of covering fewer teachers. Not only will decisions need to wait another year for a given teacher, but there are quite a number of teachers who will not have a third year of data even if they remain teaching, since teachers move in and out of tested grades and subjects.

Q: Is there a sense of how much variation over time is outside a teacher's control? There are several examples about the change in the distribution of teachers that makes it really tough to make these comparisons.

A: Yes. There is a nice study that tries to separate the variation that we see across teachers. This study separates persistent effects of a teacher from the variation in effects across years. It then further separates the variation in effects across years into sampling error and other sources of variation which includes true differences in effectiveness over time. The study finds that 30 to 60 percent of a teacher's variation over time is due to sampling error. After removing sampling error, about half of an elementary school teacher's performance in a given year is due to their persistent component and about half is due to this error that varies from year to year; for middle school teachers about 70 percent is persistent. (see McCaffrey, Sass, Lockwood, and Mihaly, "The intertemporal variability of teacher effect estimates." *Education Finance and Policy*.)

Q: Are you aware of any mixed methods that try to do some triangulation to look at the sources of this variation in student learning? Are you aware of anything that sort of gets into the process that leads to the observed growth?

A: There is a fair amount of research exploring how teachers improve over time and the effects of professional development programs on teacher improvement. There are also some studies that have looked at factors such as the turnover of other teachers in the grade or principal turnover and their effects on performance. However, in terms of research that specifically describes the causes of variation in value-added for a given teacher over time, I think we're in an early stage. There are a couple of working papers right now that are looking, for example, at variation in how much teachers improve early in their careers and the effects of schools on this improvement. I don't think there are strong conclusions from this research yet.

Q: How concerned should implementers be about the vertical linking across grades, particularly looking at issues in the sciences where there isn't a clearly articulated progression, but also for other subjects?

A: Implementers should be concerned with comparing value-added measures for teachers who are teaching different courses (e.g. different math courses that student take in a particular grade in high school). The tests that we use to create value-added measures may not measure the content of one course as well as they measure the content of another course. This is not the topic of today, but the next round of briefs in this series will include a more detailed discussion of high school value-added estimates. I am less concerned about the vertical linkage across grades. Even in elementary mathematics, vertical linkages can be questionable. Good controls for prior skills with prior test scores in multiple subjects may be sufficient controls, though I don't know of research that has assessed this directly, by, for example, comparing value-added estimates that include prior test scores in the same subject in comparison to those with more general controls.

Q: What do you think has been the best thinking on communicating error? Specifically, would you recommend that districts try to incorporate error and provide confidence intervals or some other measure of error around value-added ratings?

A: Again, there are multiple kinds of error. One kind is a sampling error, which reflects that some students gain more than others within the class and value-added measures are essentially an average of these gains. Even in the same class, other students might have learned somewhat more or somewhat less than the students who were actually in the class. Thus, the value-added measures have this sampling error. We can communicate this type of error in the form of a confidence interval, and there may be benefits to sharing value-added estimates with confidence intervals. One benefit of sharing confidence intervals is that it emphasizes the misclassifications that occur when using value-added measures to assess individual teachers. The downside of sharing the confidence intervals is that they can suggest that value-added measures contain no information since most teachers will have overlapping confidence intervals. One option is to share the confidence intervals both around individual scores and around the average scores for groups of teachers – for example, if value-added were used to identify excellent teachers, the district could share the confidence intervals around the average score for teachers in the top quintile and the average score of other teachers, as well as those around individual teachers' scores.

It is also important to emphasize that sampling error is not the only source of error in value-added measures. The thermostat in the classroom might break on the day of the test and all the students might perform worse than they otherwise would. Other idiosyncratic factors could have affected the classroom during the year. The kinds of instability that we see from year to year in teachers' value-added is quite consistent with the kinds we see in other professions for which we can measure outcomes directly, like sales. In those occupations, just like in teaching, there are problems with using outcomes as a measure of performance. Some factor outside of the workers control may affect outcomes. Even accounting for measurement error, a teacher may look as if she was ineffective when really there was a shock to her classroom that was out of her control.

Q: This next question is about teacher growth. There's a fairly well understood learning curve for teachers and their rate of improvement over their first couple of years. How do we think about accounting for the returns in experience in teacher evaluation? For example, is it fair to compare a first year to a fourth year teacher?

A: There are average gains in value-added with experience. On average, teachers gain over the first five years of teaching, and some teachers improve more than other teachers do. This said, the differences in value-added that are associated with experience differences are relatively small. Experience does not explain a lot of the differences in value-added across teachers. It probably makes sense to compare first year teachers to first year teachers and, perhaps, second year teachers to second year teachers, but whether or not the comparison is within or across experience levels should not matter very much.