Webinar 6: What Do We Know About the Tradeoffs Associated with Teacher Misclassification in High Stakes Personnel Decisions?

Q&A with Dan Goldhaber & Susanna Loeb

May 8, 2013

Q: What do we know about the relationship between false positives and the entry of quality teachers in the profession? How much do we know about what impact evaluation system errors have on workforce composition??

Goldhaber: There's almost no research on the implications of misclassification for who enters the workforce. That is partially because, in most cases, the misclassification is in one direction. Most evaluation systems suggest that teachers are doing pretty well. There's just not a lot of variation today to assess the long-term effects on prospective teachers of teaching being made to look like a less stable occupation by the attachment of high-stakes consequences to evaluations. There are upsides and downsides to the attachment of stakes to evaluations for people considering entering the profession. You can make theoretical arguments as to what would happen on both sides of the equation. Some would suggest that the instability might lead people to be less inclined to become teachers. On the other hand, the degree to which stakes reward effective teachers, teaching might look more attractive for people who think that they're going to be talented when they enter the occupation. The answer to your question is, there's really very little work that relies on observational data. There's some speculative evidence. I've read a paper by Jesse Rothstein that tries to think about this issue a little bit in a more theoretical way using assumed parameters. It's better than speculation, but it's not based on actual policy variation.

Q: Does this take on tradeoffs take into account that a low performing teacher can improve with experience? Is there a guarantee that a replacement teacher will be any better than a teacher who started off low, but for whom you have some expectation for improvement? How much do we know about how well low-performing teachers will do in the future?

Goldhaber: I think that inherent in the question is the idea that very low performing teachers may be removed. That is, removing low performing teachers might be a mistake because they improve over time with experience. So, I point out two things. One, the table Susanna talked about in her presentation takes a cohort of new teachers and looks at them in the future. So, it's inherently looking at teachers who have a similar experience level. The other thing is, there's a lot of heterogeneity in how much teachers improve with experience. So, we know that, on average, teachers tend to get better, but some teachers get better with experience and some teachers don't. If you value a student test score outcome, then I'd argue that value-added does a better job of predicting student test scores than other measures like teacher experience. I'll leave it at that and say, Susanna, please chime in.

Loeb: I think that one of the points we make here is that if you have additional information that you can bring to bear as well as the value-added, that's a good thing. So, if you have really good reason to think that a particular teacher who isn't performing well in the beginning will perform well, then that's information that can be useful. On average, at least in the studies that we've done of one big urban district, the teachers who are in the bottom group never get to the average of initial starting teachers, in terms of how effective they are. If you're just thinking about test scores, then I don't think the worry is that you wouldn't be able to get equivalently good teachers, on average. It could have system effects that we don't know about yet, in terms of either encouraging good people to come into teaching or

discouraging teaching because of the imprecision of value-added and the fact that you're going to be making some errors even if, on average, you're going to do better as a district by using value-added.

Q: What's the impact of multiple years of data on misclassification if we're looking at, say, a tenure decision at year three or four? How much better do we get at correctly classifying with two or three value-added periods?

Goldhaber: You're less likely to make errors if you use, say, two or three years of teacher value-added performance data. I think the value of additional years actually diminishes pretty quickly, probably more quickly than people think. My recollection, and this is just off of the top of my head, is that the stability of these value-added estimates increases by about 30% when you go from one to two years worth of data, then increases by another 15% when you go from two to three, and then it basically levels off. You pick up some things by increasing the number of years of performance information that you have. I'm not suggesting that you don't want to use additional years. I'm saying that the downside to waiting to make decisions is that those teachers are affecting students in some ways while you're waiting for the additional information to come online.

Loeb: We're talking about identifying teachers, but not necessarily using that identification to get rid of them immediately. It's information to use and you can use it for different kinds of interventions that are supportive and meant for improvement, rather than just dismissal. We're really talking about the use for decision making.

Q: If dismissal only happens for the bottom 0-2% in most districts, what proportion of those would be unjust dismissals? So, we're not looking at the bottom quintile, we're looking at the bottom 2%.

Goldhaber: I've never just looked at the bottom 2%, but the bottom line is that, if you look at the very bottom of the distribution, and you really only want to dismiss the bottom 2%, you're going to make a fair number of errors. It's going to look like, and again, I'm just taking these numbers off the top of my head, 50% of dismissals are going to be in error. But, importantly, you're not going to be very wrong about the effectiveness of those who are dismissed. If you're looking at the very bottom of the distribution, many of those classified as being at the bottom 2% are not going to be the bottom 2%, but they're not going to be far from the bottom 2%. A very high proportion of them are going to be at the bottom 10%.

Q: Susanna, you gave the example of changes in the work day, and students having multiple teachers. Do we have some serious tradeoffs about the challenge of analysis? If the work becomes much more complicated, doesn't it become much more challenging to actually isolate the contribution of the teacher?

Goldhaber: I think it does, potentially. The more you have multiple adults interacting with students then you worry about attribution. But, I'd also say there's a lot of room for using information to improve the quality of the existing teacher workforce, and that's not generally covered in discussions about the new performance management systems that are coming online, and ought to be.

Loeb: The only thing I'd say is that I wouldn't want to structure the job of teaching in order to make sure we get our measurement of effectiveness right.

Q: We've got a few questions that are a little further afield from the core tested grades and subjects. One is about other school-based professionals. We may have school professionals who are contributing in those classrooms, speech language, audiologists, counselors, etc. How do we think about using these data for the evaluation of those groups of educators?

Goldhaber: I think it's really difficult to think about using formulaic ways of evaluating other kinds of professionals, or in more complex school settings. It is a good argument for having systems that rely on multiple methods and allow for more nuance. One of the things that I think is really valuable about value-added is not just the information that value-added gives you about those teachers who are in tested grades and subjects where you really can attribute students to particular teachers, but it's gotten us to think more broadly about evaluation systems for educators. Hopefully, some of the broader ways that we're thinking about evaluation systems spills-over into evaluating teachers that don't neatly fit into a value-added box.

Q: This is a different take on the non-tested space. Are there studies that look at how value-added overlaps with other notions of student learning: creativity, problem solving, citizenship, many of the other learning goals for kids. Do you know of work that looks at some notion of growth in student learning or measures of teacher effectiveness outside of the traditional standards-based assessments?

Goldhaber: I'll direct your attention to just a couple studies people might want to look at. There are studies that look that the relationship between value-added and classroom observations of teachers. There's a study by Pam Grossman, and I think Susanna's a co-author. There's a study in Cincinnati by Tom Kane and some colleagues that looks at that relationship. Then there's the MET project which tries to assess teachers in a variety of different ways: classroom observation, student perceptions, and different test outcomes. So, it looks at whether value-added is comparable across different tests. Then there's also some interesting work by Kirabo Jackson, that I think you can get off his website at Northwestern, that tries to assess broader, non-cognitive outcomes of students and the degree to which teachers contribute to those outcomes and the degree to which value-added estimates of teacher effectiveness is correlated with teachers' contributions to those broad outcomes that would broadly be considered citizenship, and other kinds of student outcomes that may be related to soft skills that are quite important for a student's future endeavors in college or the workforce.

Loeb: The only thing that I would add as a sub-part of that is that in the MET study, which is a big study of different measures of teacher performance, one of the things they found that was correlated with value-added was student assessments of the teachers. Those questions captured some of these, what I would consider to be, alternative outcomes of students..

Q: You discussed that we can improve our notion of measurement with multiple measures. What research has been done that would inform the design of weighting schemes. How do you think about the use of multiple measures?

Goldhaber: When it comes to the use of multiple measures, you really have to think about what the outcomes are that you care about. One of the MET study reports is good at explaining how you'd weight different measures depending on the outcomes that you care most about. Generally, if you care about a particular kind of outcome, then you'd use measures that are related to those outcomes. For instance, if you care mainly about student achievement on the state assessment, the evidence is pretty clear that you'd weight value-added based on the state assessment very highly. And similarly you'd weight teacher observations if you care about teaching practices. Student perception surveys of teachers in the past are highly correlated with how a teacher will do in the future on student perception surveys, same with classroom observations, same with value-added. It's the base of that measure that does a good job predicting that measure in the future.

Loeb: I'd like to add one thing to that. We think of multiple of measures as something that we have to have for all teachers. The other way to think about using multiple measures is that we can use things like value-added, or some of these other measures that are possible to collect for lots of teachers, to classify teachers but then do more careful and extensive measures that you couldn't do for everyone for those teachers to really understand what's going on and reduce the measurement error in that way. In that case, you are still trying to get their contributions to student learning, but you're trying to take into account other things that might be happening.

Goldhaber: Value-added on its own, even if that's what you ultimately care about, doesn't really tell you much information about how to improve the teacher workforce, other than, perhaps, helping you make decisions about who ought to be in the workforce. For instance, you can't give teachers very nuanced feedback from value-added on its own. If you have value-added and some other measure then perhaps you can say something to a teacher about where it looks like they might want to emphasize their practice, and what kinds of professional development they might need.

Q: Susanna, we have a direct question about your response about using value-added as a screening mechanism that would trigger more careful and extensive measures. What do you think are effective additional measures to trigger?

Loeb: Districts may already be doing this for all teachers, but thorough observations are an example. The amount of time you spend in the classrooms in order to get observations to be reliable and to ensure you've addressed teacher from different angles can be big and expensive.

Q: Given the findings you've shared here, can you cite some recommendations about effective ways to assist low performing school districts to better identify, recruit and retain the effective teachers?

Goldhaber: Unfortunately, this is another area where we don't know enough and there are some important gains to be had. I'll begin by citing a paper by Jonah Rockoff and colleagues that came out in *Education Finance and Policy* a couple of years ago that looks at the variety of the kinds of assessments that are done at a point when teachers are being considered for teaching positions. This study assesses the degree to which different applicant screening tools are correlated with their later teacher effectiveness. Frankly, I think school systems could do a lot better on this front. Some schools are in the enviable position of having lots of applicants. My observation is, even in cases where a school system has a number of applicants, they often don't ask prospective teachers to do a teaching lesson. It just

seems like what we're learning about teachers is that it's hard to gauge how they'll be in the classroom without actually seeing them in the classroom. I think there's room for better talent management on the front end, in terms of deciding who's going to end up teaching in the district. On the back end, there's a whole lot more room for talent management in the way that you structure a teacher's career path, the kind of role they have in schools, and the possibility of selective retention. All of that relies on differentiating teachers. If you've got performance measurement systems that say teachers are all the same, it's very hard to use those systems for any kind of talent management once teachers are in the workforce.

Loeb: In many schools, even when the leadership really knows who the most effective teachers are, they often do little to tell those teachers that. A lot can be done, particularly in the retention of effective teachers, by simply letting them know that they are recognized for what they are doing and by encouraging them to stay. See what it would take to convince them to stay. Some analysis that we did seemed to suggest that the most important factor across schools, in terms of retention, was not getting rid of the least effective teachers, but in their ability to keep their most effective teachers. Most teachers who leave report very little encouragement to stay.

Q: This is a value-added compared to the status quo question. How would you characterize the misclassification risks associated with value-added compared to what we're currently using for high stakes decisions, seniority or credentialing? Is it a really large improvement to move to value-added?

Goldhaber: If the outcome you care about is student test scores, I'd characterize it as an incredible improvement over using just something like seniority to make decisions about layoffs, for example. Both Susanna and I have done work on layoffs where we show that you get very different teachers laid off if you're using a value-added metric as opposed to a seniority metric. That presumes that test scores are the thing that you privilege most. I want to emphasize, it's not even close. This came up in an early question about seniority. The bottom line is, teachers get better as they gain experience in their careers, but there's a lot of overlap in the distribution of effectiveness between first year and third year teachers. So, you are probably better off having a teacher at the top of the distribution of novices than at the bottom of the distribution of third year teachers.

Loeb: That probably is an understatement. You might be better off at the 50th percentile in one and the 40th in the other. The differences across experience levels explains little of the differences across teachers.

Q: What hope do you have for the improvement of measures of teaching? How, when, and by whom will measures of teaching effectiveness be improved?

Goldhaber: I've heard people talk about the improvement in value-added. I think we're going to get very little improvement in the predictive power of value-added estimates. So, my hope is that we get improvement from other measures. I'm cautiously optimistic. The bottom line is, this is a cultural question to me, or a political question about the degree to which professionals in schools are able to cut against what is now the norm of rating all teachers to be the same. So long as that continues to be the norm and, unfortunately, I would say there's suggested evidence that even the new teacher evaluation systems don't look terribly different from the old ones. As long as that's the norm, we're not going to get

very far. I think there's evidence that when you ask principals how effective their teachers are, they do a pretty good job. I think that some of that information resides in the school building, even if it isn't showing up in what are now the formal evaluations.

Loeb: I would agree with Dan. I'm less optimistic about the measures, but I also think they're less important. I am optimistic about the better use of information for decision making. It will not necessarily be by formula with value-added at the district level. I think we can do a lot to make better use of information for local decision making. Value-added is just part of what goes into those better decisions.