Q&A with Stephen Raudenbush
September 11, 2013

Q: A lot of LEAs now use multiple measures, including value-added, in teacher evaluation. Does this practice mitigate these potential biases you mention? Could it exacerbate them?

A: I think it may conceivably help. Let me just tell a little story about some analysis I did using the Tripod survey, which is a study of student perceptions of quality developed by Ron Ferguson and scholars at Harvard. We did a comparison using Tripod, and we found that the rankings within and between schools were very similar on Tripod. We think the reason is that students are comparing the teachers they're writing about to other teachers in the same school. We find that there isn't much variability between schools on Tripod, it's mostly within schools. For student perceptions, the problem that we're describing here doesn't seem to be as big a problem as it does for test scores. It may be that on other measures, observational methods, perceptions or others, it may be that this isn't such a big problem. We need to do the research to prove it, but I do have the evidence, at least with respect to student perceptions. Student perceptions are predictive of future learning of students that that teacher encounters. That's been an interesting result from the Measurement of Effective Teaching Project.

Q: From the point of additional measures, do student growth percentiles eliminate these problems that you've mentioned, or do they fall under the same problems of value-added growth modeling?

A: I think the student growth percentile is an attempt to get at this a little because each student, according to that methodology, is going to be compared to students that are similar. It asks whether a child is doing better than the children in the same range of achievement. In that sense, it means that it's helping, perhaps. The problem is that it's the collection of students that the teacher faces that really is still a challenge. The collection of students makes a difference – that is, it makes a difference whether a low achieving student is in a classroom of all low achieving students or not. Also, it makes a difference whether you have remedial resources for that child. So, I don't think this problem is solved by the growth percentile methodology. It's, perhaps, ameliorated somewhat.

Q: Is there any validity in using value-added to evaluate different types of school professionals? So, people who perform different jobs other than teaching. What does the research tell us on that?

A: I don't know of anything like that, except to say that there is research literature on school level value-added, which could be attributable to the entire school staff. I wouldn't see how it would be specifically attributable to any person who's doing anything in particular, such as counselors. I haven't seen any research of that type.

Q: You mentioned that value-added confounds teacher and school effectiveness. What advice would you give a district in terms of other variables that might help mitigate this confounding?

A: You can't really control that away in a value-added analysis. Here's the problem. Let's say I had some information on which schools are more effective. The problem is that the teacher effectiveness is correlated with the school effectiveness, and also who the students serve. It's just not easy to solve that problem. If I had good evidence on how effective schools were, I would tend to stratify the schools into subgroups that are similarly effective and compare teachers working within the subgroups.

Q: Is there a difference with regard to these issues between the two-stage and one-stage value-added model?

A: If by a two-stage model we mean computing predictions for every student and averaging the residuals and then assigning those to the teachers, there is a difference. I'm not a fan of that approach, by the way, because I think it's very difficult to interpret those results. But, those results are not invulnerable to this critique. I mean, you still have the same problem whether you do a one-stage or two-stage analysis. I thinks that's what's meant by the two-stage.

Q: You mentioned that you come from an area of Chicago that's highly segregated. Is this source of bias only a concern for a few highly segregated districts, or does this issue touch most districts? What percentage of districts need to worry about this?

A: That's an empirical question. The methods that I just described, which are described in more detail in my brief, can actually give you a very precise answer to that question. We can actually get data that will tell you. Essentially, it depends on how much variability there is between schools in their prior mean achievement and how important the prior mean achievement is in predicting the outcomes of the students. There's a simple formula for that. My sense is that most districts are segregated in one way or another. Chicago is famous for racial segregation, but when I lived in Lansing, it had socioeconomic segregation. Even when I lived in Ann Arbor, I noticed that different schools were serving kids of different socioeconomic levels. Residential neighborhood segregation is pretty much a fact of life in America. Now, I'm sure there are some examples where that isn't true, where you see elementary schools serving kids from all backgrounds. Maybe there are some aspects of school of choice where you don't have certain catchment areas. But, by and large, historically, school catchment areas have some degree of residential, income, or racial segregation.

Q: Do the methods you describe take ceiling effects into account for teachers who are already in the top percentiles?

A: Ceiling effects would actually exacerbate the problems that I'm describing, because ceiling effects would mean that people who were high in their pretest would have less to gain than people who were low. If we were comparing teachers who were teaching high achieving students and low achieving students we could expect the evaluation would be biased against the teachers teaching the high achieving students. They would have the expected lower growth rates. That actually is a perfect example of how you get into trouble when you're comparing students who are starting at very different places in the distribution. That's the lack of common support problem.

Q: Can you say a bit more about what you mean by an upper bound on bias from a failure of common support? What do you mean by this?

A: Imagine you're a parent and you're picking a classroom. Let's say you have a list of all the classrooms in a city. You want to pick the best possible classroom; the one that will maximize your child's learning. You might not care whether the growth is because it's a great teacher in a great school or whether it's a bunch of great peers who are really supportive, and it's a safe area. For all those reasons you gain a lot. The parent might not care. On the other hand, the superintendent of schools or the person who's doing the evaluating of teachers, they don't want to give teachers credit for just having favorable

neighborhood conditions and having high achieving peers and supportive parents. You want to reward people for their skill. The problem is that the skill of the teacher and the contextual conditions that I described are likely to be correlated, but we don't have an independent measure of the skill of the teacher. What happens is that if the contextual conditions vary a lot, we have a huge amount of bias. The logic of it is, the more the contextual conditions vary, the greater the risk of bias. That's what leads us to the measure of the upper bound on the bias. It's literally the product of the standard deviation, the variability between schools in their prior mean achievement multiplied by the effect of those contextual conditions on the outcome. It's that product that sets the upper bound. It's very sensible. The more heterogeneous schools are in who they take in and the more that matters, the bigger the problem you're going to have in isolating the effect of teacher skill.

Q: Do you have any recommendations for weighting value-added against other measures?

A: At this stage I'm more concerned that the value-added information we're feeding in is fair than how much we're weighting it. I'd like to make sure that when we compare teachers we're comparing teachers who are doing similar work with similar students in somewhat similar environments, then use those value-added. How much you want to weight that is really a policy issue that I don't feel qualified to answer because it depends on the goals of the school system and how important achievement in that domain is relative to other kinds of things you want people to learn. There may be other kinds of skills that are also important that aren't measured by achievement tests. To me, that's a policy issue. It's a substantive issue that goes beyond what I can say with statistics. My main concern is that whatever we feed into the system is fair, as much as we can.

Q: In the development of these fair measures, you mentioned that it helps to have good pretest data. Can you expand on what you mean by good pretest data that could be used as a baseline? What are its characteristics?

A: If the outcome is going to be the score on a math test and we have a reliable test that was given at the beginning of the year or, perhaps, at the end of the previous year, the correlation between the pretest and the posttest is very high. It could be .8 or even .9. By reliable test, I mean that it's a well put together test with well established reliability; it's long enough, that is, it has enough items so that we have a reasonably stable estimate of a child's ability; and it covers the various topics that we think are important. That very high correlation between the prior measure and the post-measure is what really works for us in doing statistical adjustments. The more strongly correlated the pre-measure is with the post-measure, the more effective the statistical adjustments will be. The pretest of a skill, like in math, is usually the very best in the sense of correlation measure of the posttest. Now, you can use multiple pretests. If you have a series of tests, they could be used to an advantage. The whole past record of that student's achievement may be extremely predictive of where that child is going to be at the end of the year. Therefore, that would be a really good pretest.

Q: We have a clarification question on common support. Is the requirement for common support that every classroom have some comparable students with all other classrooms, or do some classrooms need to overlap?

A: That's a great question statistically. I think the answer to that question is that it may depend on whether you think that some teachers are better at teaching some types of kids than other types of kids.

Let's say I have a reading teacher on the south side of Chicago in a low-income, all-minority neighborhood who's just great at teaching reading. If you think that person would also be great at teaching reading in a very different kind of setting, let's say a very affluent part of the city with high income kids. As long as you have some classrooms that strongly overlap, there's no classroom that has no overlap with other classrooms, you might be ok. But, I think that if you're using that kind of logic you have to be assuming this kind of variance of teacher skill across different kinds of kids. There is a Carnegie brief on that by Susanna Loeb on whether there is evidence that teachers are better with some kinds of kids than others. You might want to look at that brief because I think that would be the key. If we really think that a good teacher is a good teacher, then we wouldn't have to have common support with every other classroom. That's my thought.

Q: The next question is about applying this concept of common support to other types of measures. We know that many districts are using value-added to evaluate teachers as part of a more comprehensive evaluation using other types of measures. Can we apply this concept of common support to other measures? Can other measures (teacher observation, student surveys, etc.) be fair in the absence of common support?

A: That's a great question. I think the answer is that it's always a concern. If I want to compare two teachers, or two curricula, or two therapists, two anything, and I don't randomly assign participants to those two units, then we're going to use statistical adjustment. The problem of common support is going to universally arise as a question. Whatever we try to do to use statistical adjustment, equating kids is likely to not work well if we don't have good common support. It's a very general principle. There's a whole field now developing of causal inference in social science. It's a very exciting field. Causal inferences, in the absence of random assignment, are problematic if you don't have common support on whatever background measures you're using to do statistical adjustment.

Q: We have a question on how this concept of upper bounds can be practically embedded into teacher evaluation systems. Many districts use value-added as part of teacher evaluation. How can these sensitivity checks for bias inform teachers' evaluation ratings?

A: These sensitivity checks could be used to decide whether or not you're being too global in your comparisons. By global I mean that you're comparing teachers who are teaching kids who have very different preparation and background. That makes a difference. These sensitivity checks will show you that you've got a problem. I would recommend finding a subset of teachers who teach similar kids. It's as simple as that. I'd want to check out the sensitivity. I don't know. It's a simple approach. It's not hard to compute, but a lot of districts don't have great analytical capacity to do these sensitivity checks, so it's a problem. I'd argue that we need to develop the capacity of school districts to get better at doing analysis so that they can use some of this.

Q: We have a theoretical question. Why do we need to apply value-added to the whole district in the first place?

A: That's a great question. I think that, from what I've read, and I'm not particularly an advocate of why you have to do it, one argument would be that if you want to give the district greater statistical power in making personnel decisions they would have this information on all the teachers in the district. If they believe that value-added is fair, then they can say that the district has a fair procedure for evaluating

teachers. To some extent I do think that the district-wide approach is predicated on the idea that the district will have more control over personnel decisions. It's something we could debate. I happen to have done some other research suggesting, perhaps, that it's better to rely on decisions made at the school level; principals, literacy coaches and instructional coordinators. People who have a lot of information about teachers at the school level might be better able to make some of these personnel decisions. That's an interested question about management. Where should the decisions be made? A well-functioning school is rich in information, whereas a big district doesn't have a lot of information about individual teachers. If that's true, maybe the district shouldn't be making high stakes decisions. It's a great thing to debate. Again, I don't want to push my view on that, but I do think some of the pressure to have every teacher in the whole district be evaluated is to give the district central office greater power in making decisions.

Q: Many people don't like value-added because it forces all teachers into one distribution of effectiveness and percentiles, as opposed to a criterion reference, that is, teachers who have a certain score of effectiveness. What is your response to this idea that all teachers could become effective over time?

A: Well, that's a great point. I'm very intrigued by some of these emerging formative assessment systems that tell you where you're kids are. Are they on-track, and are they moving in the right direction? We have one here at the University of Chicago. In charter schools that the university runs, my colleague Tim Knowles and Tony Bryk were involved in developing this, the STEP Assessment. It tells you what step children are in their literacy. It's criterion reference. Very specifically, we know what children know and can do in their reading, and whether they are they moving up the ladder. So, if every teacher is moving kids up that ladder, that's the kind of thing you might know, probably at the school level. I'm not sure, but I think it's an extremely intriguing idea. You might also find in some cases where some teachers have kids and they're not moving up. They're not learning how to read. There may be interventions taking place and still the kids are not learning how to read. Of course you have to ask, can you help this teacher get better? Is this teacher better off doing something other than teaching? You've got to have these kids learn how to read. So, the criterion reference can also lead to some pretty hard-nosed decisions about things, but I think it's a very intriguing idea.