Q&A with Dan Goldhaber
January 9, 2013

Q: We use the Colorado Growth Model and the level of analysis is the school. Schools are sanctioned and even shut down when they don't perform well. We don't focus on individual teacher effectiveness at all. Is this an appropriate use of student growth measure?

A: I can't say what is or isn't appropriate because it really does depend on the way that a model is used. What I can do is comment a little on the school versus teacher, regardless of model. Generally speaking, when you have a bigger sample you'll have greater confidence that your estimates are, in fact, picking up something real about performance. So, you'll have a little bit better reliability at the school level than at the individual teacher level. As I began, I can't really speak to whether it's an appropriate use without knowing more about the way that it's being used. To expand on that answer a bit, echoing my sentiment about individual teachers, my take would be that if we're talking about higher-stakes uses, and shutting down schools is a very high-stakes use of a model, I would want to make sure that I had the most accurate possible assessment of what the contribution of the school is to student learning. So, if I'm looking for accurate, I'd be less inclined to use the Colorado Growth Model and more inclined to use a model that explicitly included student covariates, because a student's poverty level, for instance, seems to matter for achievement growth and is, to my mind, outside the control of schools. I believe the Colorado Growth Model now does allow for some of those kinds of covariate adjustments, so it's not necessarily a tradeoff between value-added and Colorado Growth Model. It's a question of what kind of student factors are you going to account for.

Q: Back to the tables you presented showing the correlations between different models, along the diagonal, we saw that the correlations were highest for the first and fifth quintiles. Why would correlations between different growth models be highest for teachers at the extremes?

A: It's an artifact of the fact that if you're in the top quintile you can't do better than the top, and if you're in the bottom quintile you can't do worse than the bottom. There may be some differences across models in the *absolute value* of the performance measure for teachers at the extremes of the distribution, but, for instance, a teacher that scores at the very top scores at the very top of the distribution, regardless of whether they're at the top by an inch or a mile.

Q: The MET results could be interpreted to suggest that classroom observations aren't worth the cost since they don't meaningfully add to the predictive power of teacher evaluation, and that value-added is still the best predictor. What are your thoughts on this?

A: I think it is true that if you're trying to predict achievement on standardized tests, value-added is a much better predictor than basically any other measure of teacher performance. That's true for teacher observations, student survey assessment, and other ways for assessing teachers. Now, that doesn't necessarily mean that including classroom observation doesn't improve your prediction. I believe that study suggested that it does improve it somewhat, even if slight. From a technical perspective, it may be the case that it doesn't buy you much more when you're trying to improve the prediction of future student achievement. That said, there are lots of other reasons that you might want to do classroom observations. One of the reasons is that, while most would agree that how students perform on standardized tests is an important measure of student achievement and an important measure of what

teachers are doing in the classroom, it's not the only measure. As you can imagine, some classroom practices may be important for the development of, for instance, students' soft skills, which are important for later labor market outcomes. But, these skills may not show up in achievement on a standardized test. The big caveat here is that our outcome measures standardized tests. In a sense, it's a little bit of an unfair race between classroom observation and value-added in predicting future test scores because value-added is likely to be so aligned with test score improvement. The second thing that I think is quite important to think about is the fact that if all you do is have value-added, you can only really use that measure for some purposes. It doesn't give you much of a handle on what kind of feedback to give teachers about their practices and ways they might improve practice. This is a good argument for a balanced approach to teacher performance assessment.

Q: Our student population is relatively homogeneous. Will covariates matter?

A: Covariates are less likely to matter the more homogeneous a student population is, in particular the more that the student population is equally distributed across classrooms in terms of demographics. In the extreme, if students were randomly assigned to teachers, then you wouldn't expect to see much if any difference between a student covariate value-added model with one that didn't include student covariates. But we know from a lot of evidence that, in general, students are nowhere near randomly distributed across teachers. For instance, more junior teachers are typically teaching kids who tend to come to class less academically prepared. The short answer is, the more that students are equitably distributed across classrooms, the less the covariates matter.

Q: School-wide value-added measures are being used as part of evaluations for non-tested teachers. What are your thoughts on that?

A: Ultimately, when we talk about impact, we care about what the implementation of a new evaluation system will mean for student achievement. We can't assess that until we get new evaluation systems implemented because the impact will depend in large part on how teachers respond to new systems. So, the short answer is that we don't know. If the question about impact is what will it do to the ranking of teachers in non-tested subjects in a school where value-added from the tested subjects is applied to them, then it comes back to these same issues about modeling and do the models have covariate adjustments or not, and are the non-tested teachers in schools that are advantaged or not. The answer isn't fundamentally different than the answer for the teachers in the tested subjects, if I'm understanding the question correctly. If you're in an advantaged school, teachers in tested and non-tested subjects are going to look more effective with models that do not have covariate adjustments and relatively worse with models that do have covariates. The reverse is true for teachers in schools serving disadvantaged students.

Q: Do you know what the research is on the role of student attendance as a covariate? Do any of these models deal with attendance or dosage in some way that we should be concerned about?

A: Some models do deal with dosage. You have the question when a student moves from teacher A to teacher B somewhere in the course of a year -- how much do you attribute student performance to teacher A or teacher B? There's some research that folks from Mathematica have done on that issue. I think a more difficult question is, should student attendance be explicitly included in a value-added model in the same way that one might include eligibility for free and reduced price lunch. You could

argue different sides of this. I don't like the idea of including student attendance, because a student's attendance is arguably a factor that a teacher has some influence over. I could imagine cases where you might want to include a measure of a student's attendance in prior years, but you could envision that including current attendance in the year that you're measuring student test score outcomes, you could have some perverse incentive effects. For instance a teacher suspects that a student is not going to do well on a standardized test, they may well have the incentive to discourage that student from being in class that often because of the way that the attendance factor could factor into their overall score. This may sound convoluted, but it's actually true that you could get some perverse effects. So the bottom line is that I don't like the idea of including concurrent measures of attendance in a model.

Q: At a state policy level, we've seen rules around systematically excluding mobile children. That has some pretty unfortunate consequences as well, doesn't it?

A: Right. I don't know that you'd want to exclude mobile children. I'd argued that you would want to apply different business rules regarding student mobility and teacher attrition to determine what kind of rules seem sensible so that you're being fair to teachers and not crediting a teacher for a student who's not there for the majority of the year to that teacher, but also balancing that against the potential that the rules that are created mean a great majority of kids don't count toward a teacher's performance evaluation. And again, you want to consider the potential of perverse incentives because teachers do not have a direct incentive to ensure the test performance of students they know do not count toward their performance evaluation. You'd want to be transparent about what the implications of the rules are, both for teacher rankings and students covered, in the same way that you want to be transparent about the implications of model specification.

Q: It would be great to try out data with several different models, but we have contracts with vendors and they charge. What would you recommend, to some extent, that people are beholden to their budgets that may not extend to a multiple models tradeoff study?

A: I'd make two arguments. First, I think that doing robustness testing and explaining and exploring the tradeoffs I've discussed really ought to be part of any contract and really ought to be part of the discussion that a vendor has with districts or states. I think, oftentimes, those kinds of discussions do happen. It may happen behind the scenes but they often happen. I'd also say that researchers are often looking to access really interesting datasets and might be willing to do some of the modeling that I'm suggesting for free, if doing so grants them access to interesting data. So, that's often a low-cost avenue for states and districts to pursue. For instance, the work that I presented here today is based on data from North Carolina, and North Carolina is one of the states that has developed a nice system for allowing researchers to access the data. So, it's not surprising that lots of the value-added research that we've seen gain prominence in the country is based on North Carolina data, at no cost to the state.

Q: How are teacher characteristic variables (teacher education, certification, advance degrees) included in these models and do they seem to have an impact?

A: Those variables are not included in the models that I describe here. For lots of research you do include these variables because you may care, for instance, about learning something about how effective are TFA versus traditionally licensed teachers, or what is the effect of having a more experienced teacher versus a less experienced teacher. In some cases you may care about teacher effectiveness controlling for, or factoring out, some of the things that may matter. Okay, so what

matters? In general, having a Master's degree does not matter at all. In general, teachers tend to get better early on in their careers, that is, teachers typically become more productive in the first three to five years. The degree to which graduating from a traditional teacher training program matters and passing all the required licensure tests tends to vary from state to state because the licensure systems vary from state to state. Now, do you include those kinds of factors when you're doing teacher performance evaluations? Maybe. I would argue that in most cases you wouldn't want to included those factors because you want to know, taking Chris Thorn as a whole person, how effective is Chris compared to Dan? For a performance evaluation you probably don't' care whether Chris is effective because he graduated from the right school or got the right professional development, or just that Chris is a star teacher because of something very specific to who Chris is. Whatever makes him a star teacher, I want to reflect that. Having said that, I could imagine in some cases you might want to do some adjustments. For instance, you might include teacher experience in a model because you might want to say part of why Chris looks so good is because he's a fifth year teacher and Dan is a first year teacher. Adjusting for teacher experience would help you to get an apples to apples comparison, knowing that Chris has more experience under his belt.

Q: Teachers of advantaged classrooms argue that value added is unfair because their students don't have as much room to grow. Given what you've been talking about, this comparison of classrooms made up of advantaged versus disadvantaged kids, what would you recommend?

A: As it turns out, students are more likely to do better next year if they've done better this year so teachers of advantaged classrooms tend to look pretty good under many types of student growth measures. The results I presented show that if you don't account for where students come into the classroom, whatever students bring to the table ends up being attributed to the teachers. So, if you teach advantaged kids, you look better, relatively speaking. Now, that said, the degree to which the room to grow issue is really an issue can depend on the nature of the test that is being used. For instance, tests can have a ceiling making it difficult to detect gains for high achieving students, and hence, performance for teachers of high achieving students. The evidence I've seen suggests that test ceilings generally do not appear to be a big concern for most teachers. But that doesn't mean that it's not a concern for some small percentage of the teacher workforce who may be teaching very advantaged students who came into the classroom doing extraordinarily well. And similarly, some might be concerned that some tests have a floor so you can't detect what's going on at the bottom of the distribution. So those are legitimate concerns for certain small percentages of teachers in the workforce, but not for the workforce as a whole.

Q: There has been a lot in the press recently on grit or perseverance, and notions of the contribution of the student to their own effectiveness. What other measures might be valid or reliable additions to test scores as valued outcomes?

A: That's a great question. It's also a very broad question. The short answer is, I don't know that today we have credible quantitative ways to measure things like grit, in particular whether teachers influence students' grit or perseverance. I do, however, agree that these are traits that are likely to be quite important for a person's long-term prospects. This is one reason that you wouldn't want to use value-added alone in assessing teachers, it's not clear that value-added is going to pick up those kinds of characteristics about students. That said, it's not clear that value-added would *not* pick up those kinds of characteristics. Even if a state assessment is not designed to pick up the degree to which a student really has a high degree of stick-to-itiveness, performance on the assessment may still reflect this trait because

students with higher levels of the trait might perform better on standardized tests. The bottom line here it that I think we need more research on the extent to which tests, and they certainly vary somewhat, reflect some of the non-academic skills that we think it is important for students to develop.

Q: Can you share your views on the incorporation of student survey data in evaluation? What role do you think student survey data should and could play in teacher evaluation?

A: I think that the findings on student assessment of teachers from the MET study are very interesting. It seems like the student surveys are doing at least as good a job as classroom observation at picking out what's going on in a classroom. That's a very important MET finding. Given that, I think there is reason to explore these student surveys. Now, I have a worry about using student surveys in a high-stakes way, which doesn't mean it shouldn't be done but caution is warranted. The worry is that what you might find in an experiment, like MET, when there are no stakes attached to the survey, could be quite different than what you'd find when these surveys are implemented on a wide basis, implemented for several years and maybe administered across multiple classrooms. I'll give you the extreme case that I think we need to be cognizant of. It's the eighth grader who's taking the student assessment survey for the fifth or sixth time in their middle school, also, this 8th grader is filling it out knowing that it might influence his teacher's salary next year. Do we think that in that case the student assessment survey is going to be as good a predictor of what's going on in the classroom? Do we think that in that kind of case it's going to be as good as when students are taking it for the first time under experimental conditions when there are no stakes attached? My gut says probably not, but that doesn't mean don't try it, and it doesn't mean they shouldn't be used. It means we need to be cognizant of some of those things and check it out down the line.

Q: Is it correct to say that student covariates are included to account for the assumption that teacher quality differs in advantaged and disadvantaged schools? If the average quality of teaching really is lower in disadvantaged schools, won't including student covariates cover that up?

A: Not exactly, student covariates are included to account for differences in the students that are assigned to teachers. The answer to the question you raise about whether the inclusion of student covariates will cover up teacher quality across schools is actually pretty deep into the statistical weeds, but let me begin by saying that student covariates will not themselves cover up differences in teacher quality in advantaged and disadvantaged schools. Part of the reason is that most of the estimated effects associated with being a student with a particular characteristic (e.g. eligible for free/reduced price lunch) is based on within classroom variation in achievement. That said, some types of VAMs will end up covering up differences in teacher quality between schools. For instance, the models that include school fixed effects (thereby creating within school comparisons of teachers) do mask differences in teacher quality across schools because any average difference in quality is subsumed into what gets characterized as a school effect. This is reflected in my CKN brief, which shows negligible differences across advantaged and disadvantaged classrooms in the school fixed effects VAM. Unfortunately, the flip side is that if we do not include school fixed effects we may inappropriately be attributing things that really are school level influences on student outcomes (e.g. the quality of a school principal) to teachers. For more on this topic, you could check out: http://cedr.us/papers/working/CEDR%20WP%202012-6_Does%20the%20Model%20Matter.pdf. Also stay tuned: I believe this issue of how to interpret the validity of models that compare teachers across schools will be the subject of an upcoming CKN brief.

Q: What do you think are ideal sample sizes for calculating SGP and value-added models?

A: There really is no ideal sample size. We know more about teacher performance the larger the sample of students a teacher has instructed, so, for instance, we are more likely to be able to say something about teacher effectiveness with confidence if we use multiple years of student and teacher data to inform a performance measure. However, the flip side of this is that policymakers always have to make decisions with imperfect information, and waiting for sample sizes that permit high levels of confidence may mean delaying these decisions. The Great Recession, for instance, necessitated teacher layoffs; in many cases there would not have been multiple years of student data to inform decisions about which teachers were to be laid off. Similarly, there is an argument for making earlier decisions because of time to tenure laws and the fact that one might want to address ineffective teaching early so as to make sure that fewer students are exposed to it. Finally, I want to bring up the larger issue, that we want to consider the information one can derive from VAM/SGPs with varying sample sizes in light of the information that is derived from other sources of information, such as classroom observations.

So I've totally sidestepped giving a numerical answer to this question because the answer is normative. The level of appropriate confidence that is necessary to make decisions will depend on the nature of the decision and will, regardless, be in the eye of the beholder. But, if you want to learn more specifics about how changes in samples/additional years of student performance information change the confidence in teacher performance estimates, you can look here: http://cedr.us/papers/working/CEDR%20WP%202010-3_Bad%20Class%20Stability.pdf.