

STRATEGIC **DATA** PROJECT

VALUE-ADDED MEASURES

HOW AND WHY THE STRATEGIC DATA PROJECT USES
THEM TO STUDY TEACHER EFFECTIVENESS



VALUE-ADDED MEASURES

HOW AND WHY SDP USES THEM






The mission of the Strategic Data Project is to transform the use of data in education to improve student achievement. In pursuit of this mission we conduct “diagnostic” analyses with each of our partners. One set of these analyses is focused on human capital management and teacher effectiveness. We explore questions in five stages of teachers’ careers: recruitment, placement, development, evaluation, and retention. We call this group of analyses the Human Capital Diagnostic.

We designed the Human Capital Diagnostic analyses to help leaders better understand key questions about their teaching workforce. For example: in which schools and with which students are the most effective teachers placed? How long do the most effective teachers remain in the schools or district they started in? Are teachers with advanced degrees any more effective than teachers without them? We believe that clearly understanding the answers to these kinds of questions is a prerequisite to developing focused strategies for improvement.

In the first two paragraphs of this brief, we have used the

word “effective” five times. But how do we define teacher “effectiveness”? For the purposes of the Human Capital Diagnostic, we define teacher effectiveness as success in improving student achievement in core academic subjects. Value-added models, which estimate the effect individual teachers have on student learning, provide the best widely available measure of effectiveness consistent with this definition. We refer to an estimate of a teacher’s performance based on these measures as a teacher effect.

This brief explains how and why SDP uses value-added measures for our diagnostic work. We also explain how value-added measures relate to other measures of teacher effectiveness and the limitations of value-added measures. This is not a technical document; a detailed description of our model is available at http://www.gse.harvard.edu/~pfpie/pdf/CMS_Technical_Appendix_Teacher_Effects.pdf. Rather, this document is intended for leaders and decision-makers who might use the Strategic Data Project’s diagnostic analyses to inform policy and management changes in their school districts and state education agencies.

HUMAN CAPITAL DIAGNOSTIC PATHWAY: FIVE STAGES IN A TEACHER'S CAREER				
				
RECRUIT	PLACE	DEVELOP	EVALUATE	RETAIN/TURNOVER
<ul style="list-style-type: none"> Does teacher effectiveness vary by certification pathway (traditional vs. alternative)? Where do agencies recruit their most effective teachers? 	<ul style="list-style-type: none"> Where are highly effective teachers placed (which schools and with which students)? Does teacher effectiveness vary across type of school (grades, poverty level, geography)? 	<ul style="list-style-type: none"> Are teachers with advanced degrees more effective? Do teachers become more effective as they gain experience? 	<ul style="list-style-type: none"> How predictive is novice teacher effectiveness of future teacher effectiveness? 	<ul style="list-style-type: none"> Are more effective teachers retained at higher rates than less effective teachers? Are there patterns in the movement to different jobs (into/out of teaching jobs; to/from higher needs schools, etc.)?

WHAT ARE VALUE-ADDED MEASURES OF TEACHER EFFECTIVENESS?

Value-added measures are conceptually straightforward: they aim to determine how much of a student’s academic progress from one year to the next is attributable to his or her teacher, as opposed to factors outside of the teacher’s control. The process by which we accomplish this task includes four steps.

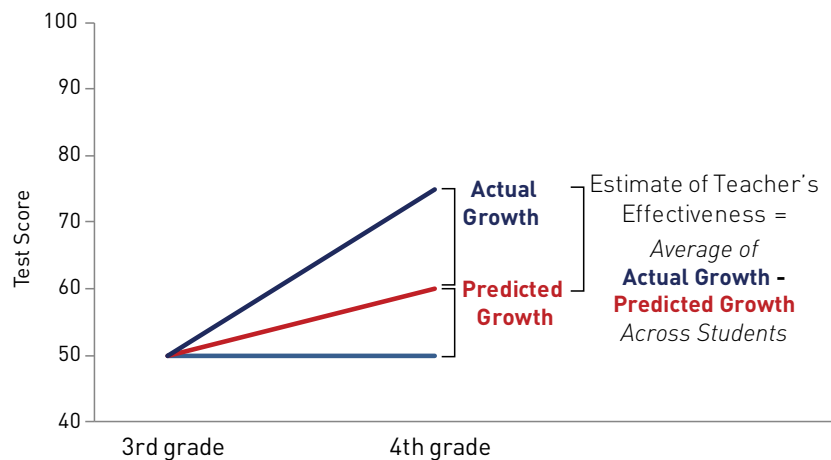
- First, we look at the prior year performance of a student on a relevant assessment. To date, we have used state assessment results due to the quality and availability of the data. However, any assessment administered to a large number of students prior to entry in a teacher’s classroom can be used as a baseline measure of performance.
- We then collect data on other student characteristics, such as eligibility for free or reduced price lunch (an indicator of economic disadvantage), and on the prior performance and other characteristics of their classroom and school peers.
- Third, using a statistical model, we use the prior performance, the current year performance, student characteristics, and classroom and school-wide factors for all students in the system to determine the performance that a typical student with similar characteristics would be expected to have on the end-of-year assessment.
- Finally, we compare the student’s actual performance to the model’s prediction for a typical student with these characteristics. We repeat this process for each student assigned to a given teacher and average the results of these comparisons to generate a value-added score, or teacher effect, for that teacher. If a teacher’s students consistently outperform the model’s predictions, this teacher is assigned a higher value-added score. If the teacher’s students underperform relative to the prediction, the teacher is assigned a lower value-added score (see figure below).

CALCULATING VALUE-ADDED	
1	Look at the prior performance of a student on an assessment.
2	Collect data on student characteristics and school-wide factors outside of the teacher’s control, such as a student’s eligibility for free or reduced price lunch.
3	Using a statistical model, use the previous performance, student characteristics, and school-wide factors to model how well a typical student with these characteristics does on the assessment at the end of the next school year.
4	Compare actual student performance to the model’s prediction. Generate a value-added measure, or teacher effect estimate, by comparing the predicted scores to the actual scores.

The statistical models used to generate value-added measures are complex, but the principles driving the model are clear. The model aims to separate out the effect the teacher has on his or her students’ learning from factors teachers have no control over such as student poverty, English-learner status, and the prior academic performance of classroom peers. Value-added technicians often refer to these variables as “control variables” in that they “control” or account for the factors outside of a teacher’s influence and help isolate the component of a student’s performance gains that can be attributed to the teacher.

HOW DOES VALUE-ADDED WORK?

Previous performance, student characteristics, and school-wide factors are used to **predict growth** at the end of the next school year for a given student. Then, the difference between the student’s **actual growth** and **predicted growth** is taken. The average of this difference across students is the value-added or effect of the teacher.



The complexity of the value-added model means it is not fully transparent to those without strong statistical backgrounds. While this may limit their usefulness for some audiences, we believe value-added measures are helpful to leaders who are focused on significantly improving teacher effectiveness.

WHY DO WE USE VALUE-ADDED MEASURES?

With no pun intended, we think they add value.

Straight proficiency measures (the primary yardstick for No Child Left Behind) are poor measures of teacher (or school) effectiveness. Schools (or teachers) with students who come from advantaged backgrounds or who enroll students who are already high achieving are likely to do well on tests and achieve high proficiency levels. Conversely, schools enrolling students who are significantly behind their peers will likely perform less well. But the teachers of these students might be helping them improve mightily—adding great value, in other words—while the teachers of high-performing students might bask in the glow of impressive proficiency rates...but not really help their students learn very much.

Unlike proficiency rates, value-added models attempt to account for where students start. Therefore, teachers whose students start far below proficiency might still have high value-added scores even if their students still end up with fairly low proficiency rates. Teachers should be credited for how far students travel under their guidance, not the achievement level students already have when they enter the teacher's classroom. Value-added models attempt to create a fairer comparison between teachers than simple level or proficiency models.

A second reason we use value-added measures is simply the absence of widely used alternative measures of teacher effectiveness that bear a relationship to improvement in student outcomes. The most commonly rewarded indicators of teacher quality—years of experience and advanced degrees—account for little of the variation in teachers' performance in improving student achievement.¹ Until very recently, the binary teacher evaluation systems used in the vast majority of school districts did a very poor job of differentiating teachers at all—with up to 99 percent of teachers rated as "satisfactory".²

Instead of using weak proxies for teacher effectiveness that capture little of the variation or bear no relationship to teacher effectiveness, value-added measures aim to link teacher effectiveness directly to student learning.³ As other measures of teacher practice that relate to student outcomes

are developed, they will provide useful additional views into teacher effectiveness.

Importantly, the data needed to construct value-added measures for large numbers of teachers are readily available. After ten years of No Child Left Behind, there is a terrific amount of untapped data about student achievement in the form of state assessment results. Because districts and states already capture these data, analyzing them generates no additional data collection costs but can generate great insight.

To be clear, we do not assert that state assessment scores are fully representative of a student's learning over the course of the year—the quality of state tests varies widely and even the best tests do not comprehensively assess all that we want our students to know. Nonetheless, they are the measures for which schools are held accountable and substantial evidence indicates that they are strongly predictive of educational attainment and labor-market success. Indeed, a recent study from the National Bureau of Economic Research confirms that assignment to a high value-added teacher is related to increased earnings as an adult.⁴

It is often noted that the value-added measures assigned to individual teachers vary from year to year. For example, in many models only about a third of teachers ranked in the top quartile of teachers in one year appear in the top quartile the next year. These annual fluctuations reflect a combination of measurement error and actual variation in teachers' performances over time. But, as a recent Brookings Institution publication points out, it is helpful to look at other sectors in which other, frequently imperfect, performance measures are used to inform important decisions. For instance, the correlation between SAT scores and college freshmen GPAs is 0.35. By comparison, the correlation of value-added from one year to the next lies between 0.30 and 0.40. Yet colleges and universities use SAT scores as a major component of their admission criteria. Similarly, the correlation between seasons for professional baseball players' batting averages is 0.36; but most general managers would not ignore players' batting averages when building a roster.⁵

Is value-added a complete measure for teacher effectiveness? Far from it. Is it available and does it add new information beyond what existing measures provide? Yes—and we should extract any useful information we can out of it.

HOW DO WE USE VALUE-ADDED MEASURES?

We use value-added measures as a way to invite deeper inquiry: what leads to a district's success in retaining high value-added teachers? What hiring and placement practices cause low-income students to have a higher probability of having low value-added teachers than other students?

We do not use value-added measures to evaluate individual teachers. In fact, we do not provide our partners with any "rankings" for individual teachers. We instead examine and share average findings for groups of teachers, often by dividing teachers into quartiles of effectiveness, but we never report any specific teachers in these quartiles.

Where fitting, we employ "pooled estimates" of value-added, meaning that we use data from multiple years for a teacher and average their estimated value-added in each year. For teachers who have been in a district or agency for more than one year, these pooled estimates greatly reduce the influence of random measurement error.

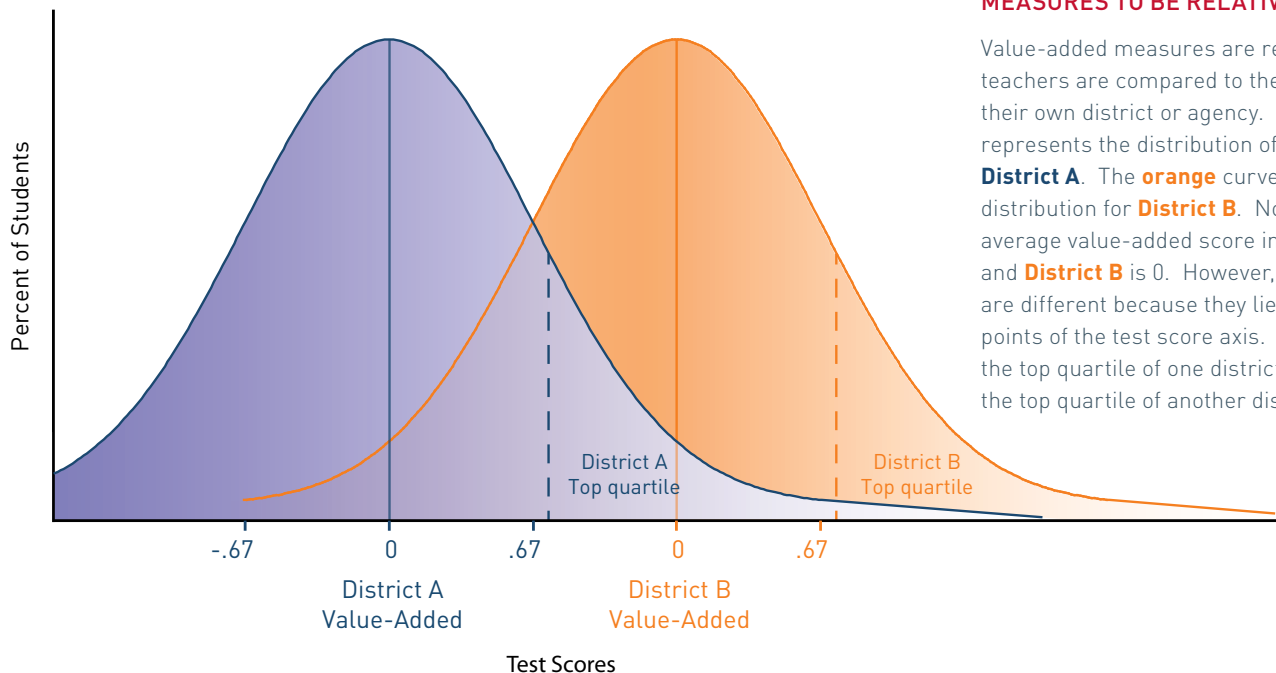
Value-added measures can be reported in a few different ways. We often convey them in percentiles. For instance, a teacher in the 50th percentile is a teacher that produces gains greater than 50 percent of the other teachers evaluated at the same time. A teacher in the 90th percentile produces gains greater than 90 percent of the other teachers evaluated at the same time.

Importantly, value-added measures are relative. That is, we are comparing teachers to the average in their own district or agency, not to an external norm or benchmark. This means that a teacher who is in the top quartile in one district, such as District A below, may not be in the top quartile in another district, such as District B. Why is this? For one, districts may have different levels of average effectiveness in their teaching force. Second, different districts may use different tests that measure slightly different skills.

HOW CAN VALUE-ADDED MEASURES FOR GROUPS OF TEACHERS BE USED?

Our mission is to educate and inform agency leaders about patterns and trends of teacher effectiveness to inform system-level management and policy decisions that improve teacher effectiveness throughout an entire agency, not to evaluate individual teachers. Incorporating value-added measures into our work has yielded a number of important insights.

For example, one trend we have found is that teachers in some of our partner districts who are hired late are less effective, on average, than their peers who are hired during the regular recruiting season. With this knowledge, school leaders could establish policy and practices to hire teachers earlier to avoid appointing less effective latecomers.



WHAT DOES IT MEAN FOR VALUE-ADDED MEASURES TO BE RELATIVE?

Value-added measures are relative because teachers are compared to the average in their own district or agency. The **blue** curve represents the distribution of value-added in **District A**. The **orange** curve represents the distribution for **District B**. Notice how the average value-added score in both **District A** and **District B** is 0. However, these averages are different because they lie on different points of the test score axis. A teacher in the top quartile of one district may not be in the top quartile of another district.

USING VALUE-ADDED MEASURES OF TEACHER EFFECTIVENESS IN PRACTICE

Dr. Lewis, the superintendent of Brookfield Public Schools, wants to know more about the effectiveness of the teachers at two schools: Elm Elementary or Oak Elementary. First, she looks at teacher evaluation scores. In both schools, 97 percent of teachers received “satisfactory” scores. She then compared the schools’ proficiency ratings on the state test. Elm met its target for Adequate Yearly Progress (AYP) with 70 percent of students scoring proficient in reading and math. Oak did not meet AYP, with only 45 percent of students scoring proficient in both subjects.

Dr. Lewis knew neither the teacher evaluation scores nor the proficiency scores told the whole story. She decided to turn to a value-added measure of teacher effects for a better understanding of the distribution of teacher effectiveness in each school. Dr. Lewis discovered that, compared to other teachers in the district, 75 percent of teachers at Oak were in the top quartile of value-added scores – indicating that Oak Elementary has many highly effective teachers, even though the school did not make AYP because its students may have started out further behind. Conversely, only 25 percent of teachers at Elm were in the top quartile of value-added scores, suggesting that although many students met proficiency, many teachers at Elm may not have contributed to high student achievement growth.

Seeing this variation caused Dr. Lewis to look into what might have generated the differences. How did the principals in the two different schools select teachers? What teaching practices were very prevalent at Oak that were not easily found at Elm? Was there any specific professional development underway at Oak?

Without seeing the differences in the value-added of the teachers in the two different schools, Dr. Lewis may not have known to dig deeper in the hiring, training, and retention practices at Oak.

Should her goal be to equalize the number of high value-added teachers at each school? Raise the “floor,” helping all teachers become more effective? Train principals to become more adept at recruiting and retaining more highly effective teachers? These are questions she would have to pursue with her staff, considering the repercussions of each move – and digging deeper in the data to develop a powerful strategy for change.

We have also found that novice teachers are, on average, less effective than teachers with several years of experience, and they are often disproportionately assigned to lower performing students both across schools and within schools. Many district leaders know that there are greater number of novice teachers in lower performing schools: indeed, these schools are often called “hard-to-staff schools.” But data in five SDP partner districts confirm that novice teachers are disproportionately assigned to lower performing students within schools. If a district is focused on narrowing achievement gaps and accelerating the learning of lower performing students, how might they adjust this teacher placement pattern to support that goal more clearly?

These are just two examples of how education agencies can use value-added based analyses to inform broad policy decisions across an agency.

WHAT ARE SOME LIMITATIONS OF VALUE-ADDED MEASURES OF TEACHER EFFECTIVENESS?

Of course, there are clear limitations to value-added measures, and we take these into consideration when designing and interpreting the results of our analyses.

Perhaps most obviously, standardized tests are an imperfect measure of student achievement. The tests do not nearly capture everything we value in what a student learns or what a teacher teaches. Also, using different tests can change teacher effects estimates for individual teachers, and, in some cases, the results from aggregate analyses based on them. At the same time, research suggests that other important measures of teacher effectiveness and practice are related to value-added scores (e.g. scores from teacher observations, student perceptions of the classroom environment⁶

The statistical model for value-added rests on some assumptions which are too pristine for the real world. For example, the model assumes only a teacher within a tested subject is responsible for all gains in that subject (i.e. science teachers do not influence student performance in math). While we would not want to create incentives that cause a science teacher to miss a teaching opportunity that is about a math concept, current value-added models, which attribute growth to specific teachers, are not designed to address this. Work is being done, however, to accommodate multiple teachers impacting student gains, so long as shared teaching occurs within a single tested subject and this phenomenon can be quantified.

Dynamic student grouping in which students are frequently regrouped with peers and new teachers also is problematic for value-added measures because the statistical models used to generate them require that we associate a group of students with a specific teacher for an entire year. Since we also account for the effect of a student's peers on the individual student, we must determine which students are grouped together for which classes. When students (or teachers) move around to address specific learning objectives, this movement is not captured.

Not all value-added models of teacher effectiveness are created in the same way. Consequently, they do not produce the exact same results. Careful decisions need to be made about which control variables to include and which teachers, students, and classes go into the sample. The answers to those questions can sometimes change teacher effect estimates.

Finally, measures of teacher value-added may be inaccurate due to information about students that is unaccounted for. For instance, if a classroom's performance on the state test was adversely affected by the noise at a nearby construction site, lower test scores would be attributed to the teacher, even though the teacher had no control over the noisy interruptions (these classroom level shocks do, however, get averaged out with additional years of data). Additionally, there might be critical missing information about individual students: a student may receive outside tutoring thus improving her outcomes above what the teacher contributed to, or a student may be tired on exam day, lowering his score in spite of very effective teaching. With a low enough score, this student could pull down the value-added measure for that teacher. Though all student scores are averaged for a teacher, it is conceivable that some individual instances could influence a teacher's value-added measure.

In some cases even data that is usually recorded may be missing. For example, we omit students from the sample who do not have a prior test score. These students may be more likely to be transfer students or students who received test exemptions due to special needs. If we systematically omit students with slower expected growth trajectories because they lack prior test scores, we may distort a teacher's value-added by not counting those students' growth in the teacher effect estimate. For these reasons, our diagnostic analyses focus on aggregates rather than individual teachers.

We believe that these caveats should inform the proper use of value-added measures and that value-added measures should be used with other, complementary measures of

teaching effectiveness, such as observations and student feedback, when possible. Nevertheless, they do not undermine the case for using value-added measures strategically for developing policies to improve the level and distribution of teacher effectiveness.

A FINAL WORD: HOW MIGHT DISTRICT LEADERS THINK ABOUT VALUE-ADDED MEASURES OF TEACHER EFFECTIVENESS?

Value-added measures are not a panacea for improving teacher quality or student achievement growth. However, we believe that value-added measures of teacher effectiveness are a useful measure of teacher performance. They offer a great deal more information about the distribution of teacher quality across an agency than most current teacher evaluation systems or achievement test proficiency rates, and they have the potential to shape policies that improve student achievement. Value-added can be a controversial subject. But it is also important to realize that districts stand to benefit from incorporating the best information they have available about their teachers, including value-added. When used appropriately, value-added measures can produce diagnostic analyses that have the potential to shape policies that serve students in the best way possible: by giving them access to more highly effective teachers.

¹ Staiger, D., Gordon, R., and Kane, T. "Identifying Effective Teachers Using Performance on the Job." The Hamilton Project, The Brookings Institution, (April 2006) and Rivkin, S.G., Hanushek, E.A., and Kain, J.F., "Teachers, Schools, and Academic Achievement," *Econometrica*, Vol. 73, No. 2 (March 2005), pp. 417–458.

² The 2009 New Teacher Project study, *The Widget Effect*, found that in evaluation systems with ratings of "satisfactory" and "unsatisfactory" 99 percent of teachers earned "satisfactory." In evaluation systems with more than two ratings, 94 percent of teachers received one of the top two ratings and less than one percent were rated unsatisfactory.

³ For more information about how our model isolates teacher effects please contact sdp@gse.harvard.edu.

⁴ Chetty, Raj, Friedman, John M., Rockoff, Jonah E., "The Long-Term Impacts of Teachers: Teacher Value-Added on Student Outcomes in Adulthood," NBER Working Paper 17699, December 2011.

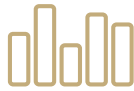
⁵ Glazerman, S., Loeb, S., Goldhaber D., Staiger, D., Raudenbush, S., and Grove Whitehurst. "Evaluating Teachers: The Important Role of Value-Added." Brown Center on Education Policy Research at Brookings. 2010 November 17.

⁶ Eric S. Taylor & John H. Tyler, 2011. "The Effect of Evaluation on Performance: Evidence from Longitudinal Student Achievement Data of Mid-career Teachers," NBER Working Papers 16877, National Bureau of Economic Research, Inc.;

Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project. Bill & Melinda Gates Foundation, 2010.;

Sartain, Laura, Stoelinga, Sara, and Eric Brown. Rethinking Teacher Evaluation in Chicago Lessons Learned from Classroom Observations, Principal-Teacher Conferences, and District Implementation, November 2011.

Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Bill and Melinda Gates Foundation, 2012.



STRATEGIC DATA PROJECT

Transforming the use of data in education to improve student achievement.

Housed at the Center for Education Policy Research at Harvard University, the Strategic Data Project partners with school districts, school networks, and state agencies in order to improve student achievement by transforming data use in education. We believe that with the right people, the right data, and the right analysis, we can significantly improve the quality of strategic policy and management decisions.

Conducting rigorous "diagnostic" analysis on teacher effectiveness is a key strategy in our theory of action. Research consistently shows that among all factors controlled by school systems, a teacher's effectiveness matters the most for student learning. Education agencies can leverage the diagnostic analyses to improve student outcomes.



Center for Education Policy Research
HARVARD UNIVERSITY