

# Value-Added Measurement

## A Brief Overview of the Rapidly Expanding Literature



**ECONorthwest**

ECONOMICS • FINANCE • PLANNING

KOIN CENTER • SUITE 1600  
222 SW COLUMBIA STREET  
PORTLAND, OREGON 97201  
PHONE • 503-222-6060

Prepared by ECONorthwest  
for the Chalkboard Project

January 2012



Technology allows school districts to collect, store, and analyze seemingly limitless amounts of data. The rapidly emerging practice of value-added (VA) measurement matches student outcomes on standardized tests to individual teachers, groups of teachers, or schools. By measuring changes in test scores year-to-year, analysts attempt to measure the value added by an educator or collection of educators. These VA estimates can be used in various ways, such as informing teacher evaluations and even staffing and compensation decisions.<sup>1</sup>

The use of VA measurement is hotly debated and appropriately so. The practice is new and fraught with pitfalls. Critics caution that small samples (e.g., 26 in a class) render the method unreliable for the purposes of evaluation, and a leading economist has shown that administrative bias in the development of class rosters can corrupt VA results. On the other hand, studies show that VA estimates of teacher quality are highly correlated with well-structured principal evaluations. And a recently released study addresses a number of common methodological shortcomings and associates high-VA teachers with students who are more likely to enroll in college and earn more money in their early careers.

This brief provides a short summary of the strengths and weaknesses of VA models early in their implementation.

## Emerging Findings on Value-Added Assessment

High-quality teachers are key to student success, but there are important questions around how to measure teacher quality and address ineffective teaching. Most school districts use some type of teacher evaluation system, but nearly all teachers are classified as good or great, leading to the perception that truly excellent teachers are not recognized and poor performing teachers are not

identified.<sup>2</sup> Existing evaluation systems do not seem to differentiate across teachers in a meaningful way.

VA measurements are based on the assumption that a teacher or school's effect on student test-score gains over time can be isolated and measured. There are several different VA models, but most compare students' actual test-score gains during the school year to expected gains, adjusting for differences in student characteristics and school-wide factors. Teachers and schools with high VA scores are those whose students' actual performance exceeds their expected performance.

Many researchers have designed studies to try to determine whether VA data are valid, reliable, and useful in evaluating teachers and schools. The primary question that fuels the debate is whether VA results are biased as a result of non-random student sorting (i.e., assigning students with certain characteristics to particular teachers). A secondary question is whether teachers with high VA scores have long-term impacts on student outcomes, or whether they are simply better at teaching to the test.

### *Teacher Value-Added Tied to Earnings and College Matriculation*

Chetty et al. (2011) provides a strong, recent case for VA reliability by developing a new way to look for bias in VA results and testing whether high-VA teachers have an effect on long-term student outcomes.<sup>3</sup> The researchers analyzed data from grades 3-8 for 1 million students from 1989-2009, linking those data with student outcomes, parent characteristics, and IRS tax records. To avoid the effects of non-random student sorting, the authors analyzed "teacher-switch" cases—where naturally occurring staffing changes resulted in a new teacher entering the classroom during the year (students began the year with one

teacher and finished the year with a different teacher). They concluded:

- VA estimates accurately capture teacher impact on student test-score gains in grades 5-8. When a high-VA teacher joins a school, test scores rise immediately in the grade taught by that teacher; when a high-VA teacher leaves, test scores fall.
- Students with high-VA teachers are more likely to attend college, earn higher salaries (see Figure 1), and save more for retirement. They are less likely to have children as teenagers.<sup>4</sup>
- Replacing a low-VA teacher with an average-VA teacher would increase an average classroom of students' lifetime income by more than \$250,000. Families would earn an average annual rate of return of 5 percent if they invested \$4,600 to give their child a teacher with a better VA score.<sup>5</sup>

The authors point out that looking back at VA data is very different than using it in an environment

where teacher evaluations and pay may be at stake; implementing VA models may change teacher and principal behavior and introduce costly personnel decisions and turnover rates.

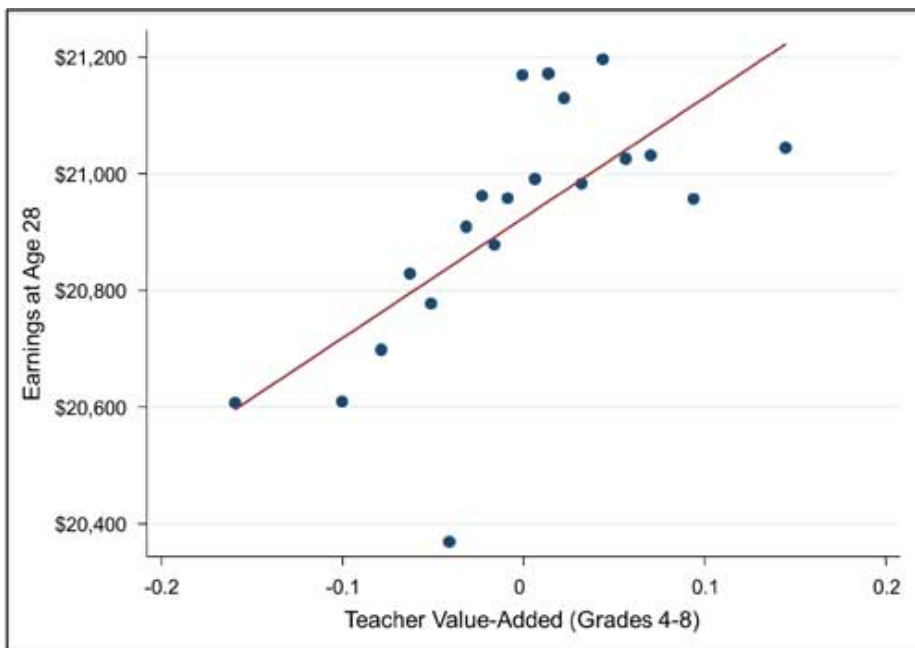
The Chetty et al. study addresses and reconciles conflicting findings from earlier papers, in large part by using teacher-switch cases to avoid biased results (see the next section on VA's limitations). Its conclusions resonate with those in Kane and Staiger (2008), who used VA results for 78 pairs of elementary teachers in Los Angeles from 2003-04 to show that VA results have strong power in predicting the next year's outcomes.<sup>6</sup> Using both non-experimental and experimental data allowed Kane and Staiger to test VA methods against a random-assignment case with 3,194 students. In this case, teacher impacts on student scores appeared to "fade out" by about 50 percent for the next two years, but they concluded that "concerns about bias in teacher VA estimates may be overstated in practice" and that VA results can produce "unbiased and reasonably accurate

predictions of the causal short-term impact of a teacher on student test scores."

#### *VA Assessments Correlate With Classroom Evaluations*

Several researchers have compared VA models with classroom observation systems, such as Danielson's Framework for Teaching (FFT). Most studies have concluded that evaluation results from well-executed classroom observations are positively correlated with student achievement scores, as measured by VA data.<sup>7</sup> For example, Kane et al. (2011) found that teachers' scores

Figure 1: Effect of Teacher Value-Added on Earnings at Age 28



Source: Chetty, R., Friedman, J., Rockoff, J. (2011, December). *The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood*. NBER Working Paper No. 17699.

based on principals' classroom observations identify effective teachers in the Cincinnati Public Schools and reliably predict test-score gains made by students in both math and reading.<sup>8</sup> In these studies, VA results and principal evaluations tend to reinforce each other.<sup>9</sup> The authors recommend the use of evaluation systems that use multiple measures.

### Limitations and Ongoing Concerns

VA models clearly have several important limitations; the inevitable influence of other factors prevents them from perfectly capturing individual teacher or school effects.

#### *Concerns About Assignment Bias in VA Models*

Rothstein (2009) reported that common VA models generate biased results and suffer from flawed assumptions.<sup>10</sup> His main premise was that, if VA models are valid and reliable, you should not be able to predict students' past test scores using their current or future teachers' VA scores. Using a large dataset from North Carolina (60,740 students in 868 schools), Rothstein found that 5<sup>th</sup> grade teachers do appear to have an effect on 4<sup>th</sup> graders' test-score gains, suggesting that students have at least as much effect as their teachers on their gains. Thus, he concluded that commonly used VA models are not equipped to handle the effects of sorting students across classrooms. He suggested that the evidence does not support the use of VA as a major component of evaluation practices nor as a primary basis for incentive-based pay.<sup>11</sup>

#### *Many Factors Can Influence Student Performance and VA Results*

Like many researchers, Linda Darling-Hammond, who led President Obama's education policy transition team, has stated that VA measures are valuable to the extent that they inform and validate a larger system of performance-based assessments and standards-based evaluations.<sup>12</sup>

Her concerns with VA measures for individual teachers include:<sup>13</sup>

- VA scores are error-prone and unstable (varying from year-to-year, class-to-class, test-to-test, and model-to-model), even when controlling for other factors
- Making high-stakes decisions based on VA causes teachers to focus inordinately on standardized tests, which aren't consistent in the types of learning they measure
- Teachers may seek to avoid teaching special education, English language learner, and low-income students<sup>14</sup>

Regarding the stability of VA assessment, an Economic Policy Institute report summarizes the factors that influence VA scores at the teacher level:<sup>15</sup>

- The non-random assignment of students to classrooms and schools (high-performing kids are sometimes sorted to low-performing teachers to increase the school's chance of meeting standards)
- Small classroom-level samples (particularly if a large number of students move between districts or states and cannot be tracked from year to year)
- The influence of other teachers, tutors, class size, attendance, parent characteristics, family resources, health, disabilities, neighborhood and peer influences
- Tests that are not aligned with classroom curriculum, or that do not measure actual achievements of students in the class

In addition, many grades and subjects don't use standardized tests, potentially excluding many teachers from VA models,<sup>16</sup> and it is difficult to measure the possible spillover effects between teachers and subjects. Finally, VA models have been shown to be most useful for identifying very high- and low-performing teachers (the top quintile

and the bottom quintile)—they are less reliable at classifying teachers in the middle.<sup>17</sup>

Some researchers have suggested that VA models are more reliable for schools than for individual teachers. For example, Schochet and Chiang (2010) found that error rates are lower when schools are the performance unit in VA models, and that VA estimates “are likely to be noisy using the amount of data that are typically used in practice.”<sup>18</sup> Larger datasets that cover a longer range of time provide the most reliable input information for VA models.<sup>19</sup>

### NEA Guidance on VA Models

The National Education Association (NEA) has affirmed that VA measurement “is an improvement over previous student achievement measures,” which compared average achievement measures across classrooms without controlling for expected performance or the various outside factors that influence achievement.<sup>20</sup> NEA has provided research-based guidelines for state and local affiliates to use when they are being asked to consider the technical aspects of a VA system proposal. Recommended best practices fall into three categories:<sup>21</sup>

#### *Accuracy and Reliability*

- Model makes use of prior information about student achievement and student, classroom, and school characteristics
- VA estimates are calculated using at least a 3-year rolling average
- VA scores are based on a minimum class size of 15 students
- Data systems uniquely link teachers, students, and subject-specific test results
- Student mobility and absenteeism are addressed as necessary

- VA scores are based on developmentally appropriate tests that are in alignment with state/district curricula

#### *Fairness*

- VA estimates are reported using a scale with interpretable and fair cut points, with input from a variety of stakeholders
- System has clearly articulated goals; evaluations and changes are based on intended and unintended consequences
- System includes a fair and feasible solution for teachers of untested grades and subjects
- System includes a fair and feasible solution for non-teaching personnel
- Detailed documentation allows the VA scores to be replicated, and external evaluations are allowed
- User-friendly materials for teachers and parents describe the system in detail
- VA scores are used as part of formative assessments and to guide professional development

#### *Feasibility*

- Amount of time required to analyze and turn around the data is clear and reasonable
- VA contractor is familiar with the strengths and limitations of the proposed VA methods (and who else is using them)
- VA contractor has extensive and relevant experience

### Summary

Everyone agrees that students deserve good teachers and schools, and that school districts should recruit and retain teachers with the training and characteristics to teach effectively. VA models provide a new and potentially valuable way to measure teacher and school impacts on

student performance, but the approach is not a silver-bullet solution to evaluation challenges.

The consensus in the research community is that VA estimates should be used with caution as part of comprehensive evaluation frameworks. Many studies suggest that VA can provide meaningful, if imperfect, information about teacher effectiveness, but more research is needed on how best to use VA information in personnel and compensation decisions. To the extent possible, VA methods should control for the effects of non-random student sorting, student characteristics, and past achievement. And VA data alone should not determine pay structures or bonuses.

VA estimates can supplement and support other measurement tools such as observations or recordings of classroom practice; teacher interviews; surveys of parents, students, and peers; portfolio reviews; lesson plan analyses; and samples of student work. Evaluation frameworks need to include multiple tools to accurately assess as many teachers and staff as possible.

The authors of a Brookings Institution report make four final points to consider about VA measures:<sup>22</sup>

- VA data and how those data are used are two different things. For example, evaluation systems can incorporate VA information without releasing it publicly or using it as the primary metric for hiring, firing, or compensation decisions. “Objectionable personnel policies” shouldn’t be confused with VA information itself.
- VA results will sometimes be wrong, classifying effective teachers as ineffective (false negatives) and ineffective teachers as effective (false positives).<sup>23</sup> Concern with the effect of false negatives on teachers needs to be balanced with a concern for the effect of false positives on students.
- VA-type measures used for high-stakes decisions in other fields (e.g., healthcare)

have similar levels of classification errors and year-to-year correlation errors as VA measures for teachers.<sup>24</sup> “VA evaluations are as reliable as those used for high stakes decisions in many other fields.”

- VA models are just as useful as other existing methods of classifying teachers. If student test-score gains are the desired outcome, VA data are probably better than other measurement methods. Ignoring VA information may lower the reliability of personnel decisions about teachers.

Teacher performance measures need to be improved, and VA estimates can play a role in those improved measures. Research, debate, and discussion around VA models should continue to be focused on their validity, reliability, and usefulness.<sup>25</sup>

---

<sup>1</sup> Examples of controversial uses of VA measures include the public release of names and scores for 12,000 teachers in Los Angeles in 2010, and the use of VA results to account for 50 percent of teacher evaluation scores in the District of Columbia school system.

<sup>2</sup> Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The Widget Effect: Our National Failure to Acknowledge an Act on Differences in Teacher Effectiveness*. New York, NY: The New Teacher Project.

<sup>3</sup> Chetty, R., Friedman, J., Rockoff, J. (2011, December). *The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood*. NBER Working Paper No. 17699.

<sup>4</sup> These results provide additional data points to consider alongside findings from Kane and Staiger (2008) and Rothstein (2009) that an individual teacher’s effect on a student’s test score performance fades out very quickly in subsequent years.

<sup>5</sup> Another study (Hanushek, E. (2010, December). *The Economic Value of Higher Teacher Quality*. NBER Working Paper No. 16606) estimates that “a teacher one standard deviation above the mean effectiveness annually generates marginal gains of over \$400,000 in present value of student future earnings with a class size of 20 and proportionately higher with larger class sizes.”

<sup>6</sup> Kane, T., & Staiger, D. (2008). *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation*. NBER Working Paper No. 14607.

<sup>7</sup> See Heneman, H., Milanowski, A., Kimball, S., & Odden, A. (2006). *Standards-based Teacher Evaluation as a Foundation for Knowledge- and Skill-based Pay*. Philadelphia, PA: Consortium for Policy Research in Education.

<sup>8</sup> Kane, T., Taylor, E., Tyler, J., & Wooten, A. (2011). "Identifying Effective Classroom Practices Using Student Achievement Data." *Journal of Human Resources*, 46(3): 587-613.

<sup>9</sup> See also Jacob, B., & Lefgren, L. (2005.) *Principals as Agents: Subjective Performance Measurement in Education*. NBER Working Paper No. 11463.

Jacob, B., & Lefgren, L. (2008). "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education." *Journal of Labor Economics*, 26, 101-136.

<sup>10</sup> Rothstein, J. (2009, May). *Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement*. Princeton University and NBER.

<sup>11</sup> See also Rothstein, J. (2011, January). Review of *Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project*. Boulder, CO: National Education Policy Center.

<sup>12</sup> See Darling-Hammond, L. (2010, October). *Evaluating Teacher Effectiveness: How Teacher Performance Assessments Can Measure and Improve Teaching*. Center for American Progress.

<sup>13</sup> Darling-Hammond, L. (2011, May 24). "Assessing a Teacher's Value: Too Unreliable." *New York Times*.

Darling-Hammond, L. (2011, May 30). "Testing Students to Grade Teachers: A Dangerous Obsession." *New York Times*.

Haertel, E., Rothstein, J., Amrein-Beardsley, A., & Darling-Hammond, L. (2011, September 14). *Getting Teacher Evaluation Right: A Challenge For Policy Makers*. Capitol Hill Briefing, American Education Research Association and National Academy of Education.

<sup>14</sup> Even when statistical controls are used for student demographic factors and school differences, teachers have been found to receive relatively lower VA scores when they teach new English learners, special education students, and low-income students.

Much of the achievement gap between affluent and poor students occurs because of summer learning loss for low-income students. See Alexander, K., Entwisle, D., & Olson, L. (2001). "Schools, Achievement, and Inequality: A Seasonal Perspective." *Educational Evaluation and Policy Analysis*, 23(2), 171-191.

<sup>15</sup> Baker, E., Barton, P., Darling-Hammond, L., et al. (2010). *Problems With the Use of Student Test Scores to Evaluate Teachers*. Briefing Paper #278. Washington D.C.: Economic Policy Institute.

The authors also note that VA assessment can result in a number of unintended negative effects, including disincentives for teachers to work with the neediest students (because test-score gains are potentially more difficult to

achieve); less teacher collaboration, increased teacher demoralization, and teaching to the test; and discouraging good teachers from entering and staying in the system.

<sup>16</sup> Prince, C., Schuermann, P., Guthrie, J., et al. (2008). *The Other 69 Percent: Fairly Rewarding the Performance of Teachers of Non-tested Subjects and Grades*. Washington, DC: Center for Educator Compensation Reform, U.S. Department of Education, Office of Elementary and Secondary Education.

<sup>17</sup> Little, O. (2009). *Teacher Evaluation Systems: The Window for Opportunity and Reform*. Research Department, National Education Association.

<sup>18</sup> Schochet, P., & Chiang, H. (2010, July). *Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains* (NCEE 2010-4004). Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

<sup>19</sup> Lipscomb, S., Teh, B., Gill, B., et al. (2010). *Teacher and Principal Value-Added: Research Findings and Implementation Practices*. Mathematica Policy Research for Team Pennsylvania Foundation.

Haertel, E., Rothstein, J., Amrein-Beardsley, A., & Darling-Hammond, L. (2011, September 14). *Getting Teacher Evaluation Right: A Challenge For Policy Makers*. Capitol Hill Briefing, American Education Research Association and National Academy of Education.

<sup>20</sup> Little, O. (2009). *Teacher Evaluation Systems: The Window for Opportunity and Reform*. Research Department, National Education Association.

<sup>21</sup> Baer, J., Gulemetova, M., & Prince, C. *Research Guidance to State Affiliates on Value-Added Teacher Evaluation Systems*. Research Department, National Education Association.

NEA has also provided criteria for evaluation and accountability methods in general: National Education Association. "New Policy Statement on Teacher Evaluation and Accountability," <http://www.nea.org/grants/46326.htm>

<sup>22</sup> Glazerman, S., Loeb, S., Goldhaber, D., et al. (2010, November 17). *Evaluating Teachers: The Important Role of Value-Added*. Washington D.C.: The Brown Center on Education Policy at Brookings.

Glazerman, S., Goldhaber, D., Loeb, S., et al. (2010, December 15). "Value-Added: It's Not Perfect, But It Makes Sense." *Education Week*.

<sup>23</sup> See, for example, Schochet, P., & Chiang, H. (2010). *Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains* (NCEE 2010-4004). Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

<sup>24</sup> For one study of year-to-year variability, see McCaffrey, D., Sass, T., Lockwood, J., & Mihaly, K. (2009). "The Intertemporal Variability of Teacher Effect Estimates." *Education Finance and Policy*, 4(4): 572-606.

---

See also Goldhaber, D., & Hansen, M. (2008, November). *Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions*. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research, Urban Institute.

<sup>25</sup> Gabriel, R., & Lester, J. (2010, December 15). "Public Displays of Teacher Effectiveness." *Education Week*.