A Thorough Analysis of Value-Added Scoring of Teachers

In this important 28-page paper published by Educational Testing Service, Edward Haertel (Stanford University) assesses the reliability and validity of using student test scores to evaluate teachers. He begins by summarizing Americans' opinions on public education: we're falling behind other nations in test scores and college completion rates; teacher evaluation in most school districts is ineffective; we're not identifying and dismissing our poorest-performing teachers; and teachers' credentials alone say very little about instructional quality. "What could be more reasonable, then, than looking at students' test scores to determine whether or not their teachers are doing a good job?" asks Haertel. "The teacher's job is to teach. Student test scores measure learning. If teachers are teaching, students should learn and test scores should go up. If they are teaching well, scores should go up a lot. If test scores are not moving, then teachers should be held accountable." A new generation of sophisticated value-added teacher evaluation seems to make this possible.

How much does teacher effectiveness matter to student achievement? Haertel asks. Researchers estimate that factors outside the schoolhouse – family background, neighborhood environment, peer influences, and differences in students' aptitudes for schooling – account for as much as 60 percent of the variation in students' test scores. Other factors account for another 30 percent. That leaves teaching with only 10 percent of total impact – not very much, but it's the 10 percent that is most amenable to policy intervention, hence the push to evaluate teachers more rigorously and "get rid of" the worst. But if the goal is to dramatically improve student achievement, identifying and removing low-performing teachers won't be nearly enough; as Linda Darling-Hammond puts it, "You can't fire your way to Finland."

However, the quality of teaching clearly matters, especially for disadvantaged students. "A classroom full of students with no teacher would probably not learn much," says Haertel, "– at least not much of the prescribed curriculum. But the relevant question is not between some teacher and no teacher, but rather between a good teacher in some sense and a poor teacher." When students have several consecutive years with effective teaching, they make significant progress – and when they have several years of substandard teaching, the achievement gap widens. Which brings us back to the question of how to evaluate and improve the quality of teaching across the board.

The simplistic approach of using students' end-of-year test-scores to compare and evaluate teachers has serious psychometric problems, says Haertel. First, because most tests don't have equal-interval scales, the impact of the same teacher depends on the achievement levels of his or her students, with the biggest variations occurring at the top and bottom of the achievement scale. Second, if tests aren't vertically aligned from grade to grade (and most aren't), it's unfair to compare the instructional impact of teachers working at different grade levels. Finally, some teachers have more challenging groups of students than others. Using test scores is therefore highly problematic – and has the unintended consequence of creating

disincentives to working with challenging students, teaching at certain grade levels, and even working with high-achieving students.

These are exactly the problems that value-added models (VAM) are supposed to solve. Their goal is to strip away the factors that are not under the teacher's control and leave just the achievement information that the teacher can control – the causal effect for which he or she can be held accountable. For example, the formula used to calculate the effectiveness of Los Angeles teachers (the results were published by the *LA Times* in 2010) included five data points: students' gender; test performance the previous year; English language proficiency; eligibility for Title I services; and whether the student began school in the LA Unified School District. Other value-added formulas use several years of prior test scores, attendance, suspensions, grade retentions, and several demographic factors. Having taken all these variables into account, the formulas aim to predict each student's potential test scores with *any* teacher. It's then possible to compare that hypothetical projection with what an *actual* teacher produced. That's the teacher's value-added.

But there are problems. First, different teachers have very different working conditions and it's impossible for a value-added formula to take them all into account. "School climate and resources, teacher peer support, and, of course, the additional instructional support and encouragement students receive both out of school and from other school staff," says Haertel, make some teacher's work much more challenging than others.

Second, U.S. schools are highly stratified by socioeconomic status, and it's difficult to project students' hypothetical scores without going well beyond the available data. "For this reason," says Haertel, "VAM estimates are least trustworthy when they are used to compare teachers working in very different schools or with very different student populations."

Third, the academic composition of each class influences the teacher's pacing, the level at which the material is explained, the amount of work assigned, and expectations. In addition, peer dynamics within classes can promote or disrupt learning. "Just about every teacher can recall some classes where the chemistry was right," says Haertel, "– perhaps one or two strong students always seemed to ask just the right question at just the right time to move the classroom discussion along. Most teachers can also recall some classes where things did not go so well." This means some teachers can move much more quickly through the curriculum than others, and because the grouping of students within schools is "massively nonrandom" and some schools have climates much more conducive to learning than others, comparisons among teachers are inherently unfair.

Fourth, researchers have discovered a number of statistical anomalies that cast doubt on value-added measures. For example, when calculating year-to-year student learning increments, one has to take into account summer learning loss – and there are major differences in how much students lose over the summer: disadvantaged students typically lose ground over the summer, while advantaged students gain. "Some of this difference may be accounted for in VAMs that include adjustments for demographic factors," says Haertel, "but once again, it appears likely

that value-added estimates may be biased in favor of some teachers and against others." In addition, researchers have found huge year-to-year variations in individual teachers' value-add scores – some veering from the top quintile one year to the bottom quintile the next. These variations reflect changes in the students taught and changes in teachers' performance – but how much from which? About "half of the variation in these value-added estimates is signal," says Haertel, "and the remainder is noise… Sorting teachers according to single year value-added scores is sorting mostly on noise." Using three years of data increases reliability, but only to about .56, which is hardly impressive.

Fifth, research comparing teachers' value-added scores with administrators' observations and student survey data (the Measures of Effective Teaching study did this) show weak correlations. In some cases, teachers who scored in the top quartile on value-added had extremely poor evaluations from observers and students. And studies comparing teachers' value-added scores using different standardized tests (in one case, the state test, the Scholastic Reading Inventory, and the Stanford Achievement Test) show huge variations. "Therefore," says Haertel, "if a school district were to reward teachers for their performance, it would identify a quite different set of teachers as the best performers depending simply on the specific reading assessment used."

Sixth, value-added scores don't take into account non-cognitive learning, which is increasingly seen as a crucial factor in students' future prospects. Some teachers are making contributions to students' life trajectories that are simply not picked up in value-added data.

Finally, value-added scores are available for only a small minority of teachers within any district.

In short, even the most sophisticated value-added model "will *not* simply reward or penalize teachers according to how well or poorly they teach," Haertel says. "They will also reward or penalize teachers according to *which students* they teach and *which schools* they teach in… Adjusting for individual students' prior test scores and other background characteristics may mitigate – but cannot eliminate – this problem."

Haertel concedes that VAM scores *do* predict important student learning outcomes. Some studies have shown long-range benefits to students who were taught by teachers with high value-added scores – but the strongest evidence comes from studies that looked at tests that were not high-stakes for teachers. He agrees that value-added data can detect real differences in many teachers' effectiveness and might be useful when researchers compare large groups of teachers to assess how well training and policy innovations are working. He says *low-stakes* value-added data could be useful to principals and teachers if used under the following conditions: (a) Scores are based on sound, appropriate student tests; (b) Comparisons are limited to homogenous teacher groups; (c) There is no fixed weight – there's flexibility to interpret value-added scores in context for individual use; (d) Users are well trained to interpret scores; and (e) Everyone has clear and accurate information about the margin of error. "These five conditions would be tough

to meet," he says, "but regardless of the challenge, if teacher value-added scores cannot be shown to be valid for a given purpose, then they should not be used for that purpose."

All that said, Haertel concludes with the following observations:

• "Teacher VAM scores should emphatically *not* be included as a substantial factor with a fixed weight in consequential teacher personnel decisions," he says. "The information they provide is simply not good enough to use in that way. It is not just that the information is noisy. Much more serious is the fact that the scores may be systematically biased *for* some teachers and *against* others, and major potential sources of bias stem from the way our school system is organized."

• "High-stakes uses of teacher VAM scores could easily have additional negative consequences for children's education," he says. "These include increased pressure to teach to the test, more competition and less cooperation among the teachers within a school, and resentment or avoidance of students who do not score well."

"Reliability and Validity of Inferences About Teachers Based on Student Test Scores" by Edward Haertel in The 14th William H. Angoff Memorial Lecture, March 22, 2013 (Educational Testing Service), http://www.ets.org/Media/Research/pdf/PICANG14.pdf