EDUCATIONAL LEADERSHIP

November 2012 | Volume **70** | Number **3**
Teacher Evaluation: What's Fair? What's Effective? Pages 65-68

All Hat and No Cattle

Jessica Holloway-Libell, Audrey Amrein-Beardsley and Clarin Collins

The saga of four teachers in Houston, Texas, shows that value-added measures are not the silver bullet some advocates claim they are.

Imagine this for a moment. A new sheriff comes to town—a real maverick—and pledges to end all crime in the county. His tactic? To purchase an intricate "value-added" formula that will measure each police officer's ability to rid his or her precinct of criminal activity. Officers who do not meet the predetermined cut scores will be terminated and replaced. Officers who "add value" will be rewarded. The trusting public endorses the new sheriff's plan.

The problems with the plan are many, however. Police officers who work in the inner city express concern that factors outside their control, such as demographics, previous crime rates, and shifting unemployment rates, will make it harder for them than for their colleagues in the suburbs to be rated "highly effective." Officers who work in the suburbs, on the other hand, complain that they will not have a reasonable chance to demonstrate improvement given the already-low percentages of crime in their precincts.

As time goes by, the police officers note that many precincts' crime rates fluctuate dramatically from year to year, making accurate determinations of performance nearly impossible. Nonetheless, determinations continue to be made, as promised to the public. Other officers note that during some of the periods when their precincts' crime rates increased, the sheriff, unaware of this information, recognized them for their excellent work on other measures. The sheriff dismisses even his own judgment as subjective.

For every concern expressed by the officers, in fact, the sheriff has a counterclaim. The expert statisticians who designed the model, he insists, have built in statistical controls for every problem. The officers are advised to trust the mathematics behind the model—although even the sheriff himself does not quite understand it.

The sheriff, to the officers' chagrin, continues to walk tall. The public, still blind to the problems, continues to back his plan. The officers are paid with taxpayer funds to keep the peace, so it only makes sense that the sheriff and the public should hold the officers accountable for doing so, even if insiders know that the new system is grossly flawed.

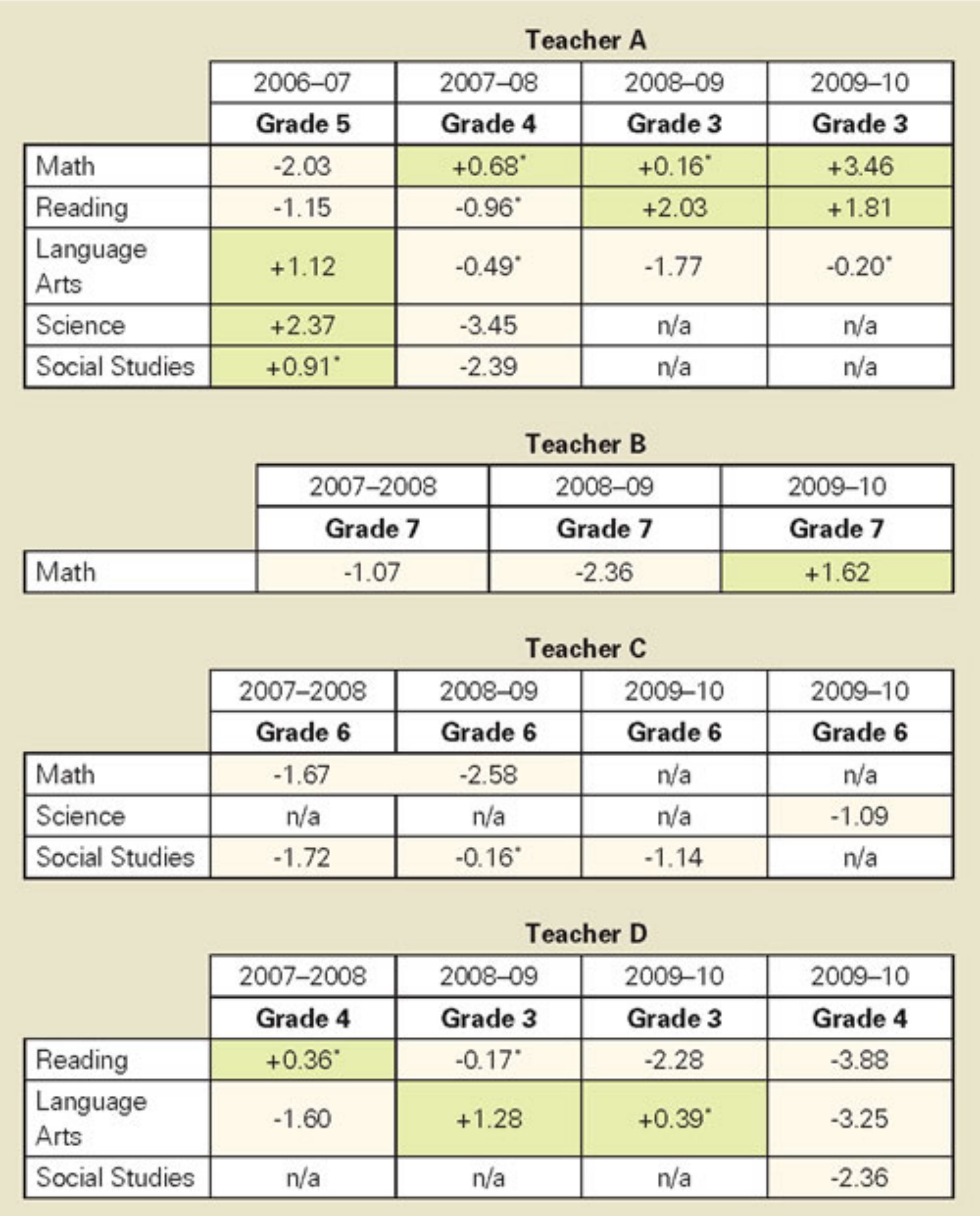
Everything's Bigger in Texas

Now imagine the following scenario playing out in Houston, Texas. This time, the sheriff is the superintendent of the Houston Independent School District, who aligns his plan for education reform with the wishes of his commander—U.S. Secretary of Education Arne Duncan (U.S. Department of Education, 2009). His plan is to reward or penalize teachers depending on whether they demonstrate that they "add value" to their students' learning and achievement—as determined by using a complicated statistical formula that shows whether their students' standardized test scores are higher or lower than predicted.

The superintendent has signed a five-year contract (at $500,000 per year) with SAS, a software corporation, to apply its Education Value-Added Assessment System (EVAAS) to judge Houston teachers' performance. This system, if "effectively implemented … allows educators to recognize progress and growth over time, and provides a clear path to achieve the U.S. goal to lead the world in college completion by the year 2020" (SAS, 2012).

Recently, two of us (Amrein-Beardsley & Collins, 2012) studied the impact of this model in Houston. We examined the cases of four teachers who were terminated in summer 2011, at least in part because of their subpar EVAAS scores (see fig. 1, p. 69). Talking to these and other Houston teachers, we uncovered some unintended consequences of EVAAS's implementation.

FIGURE 1. EVAAS (Value-Added) Scores of Four Teachers Terminated from Houston Independent School District in Summer 2011



Teacher A

	2006–07	2007–08	2008–09	2009–10
	Grade 5	**Grade 4**	**Grade 3**	**Grade 3**
Math	-2.03	+0.68*	+0.16*	+3.46
Reading	-1.15	-0.96*	+2.03	+1.81
Language Arts	+1.12	-0.49*	-1.77	-0.20*
Science	+2.37	-3.45	n/a	n/a
Social Studies	+0.91*	-2.39	n/a	n/a

Teacher B

	2007–2008	2008–09	2009–10
	Grade 7	**Grade 7**	**Grade 7**
Math	-1.07	-2.36	+1.62

Teacher C

	2007–2008	2008–09	2009–10	2009–10
	Grade 6	**Grade 6**	**Grade 6**	**Grade 6**
Math	-1.67	-2.58	n/a	n/a
Science	n/a	n/a	n/a	-1.09
Social Studies	-1.72	-0.16*	-1.14	n/a

Teacher D

	2007–2008	2008–09	2009–10	2009–10
	Grade 4	**Grade 3**	**Grade 3**	**Grade 4**
Reading	+0.36*	-0.17*	-2.28	-3.88
Language Arts	-1.60	+1.28	+0.39*	-3.25
Social Studies	n/a	n/a	n/a	-2.36

Scores shaded as green indicate that the teacher added value according to EVAAS data in comparison with similar teachers across the district. Scores shaded as pale yellow indicate the opposite.

* Indicates that a score was not detectibly different from the reference gain scores of other teachers across the district within one standard error. These scores, however, are still reported to both the teachers and their supervisors for decision making.

Four Terminated Houston Teachers

All four of the teachers in question were female elementary school teachers from racial minority backgrounds. Two were traditionally certified, and two were alternatively certified. They had logged an average of 11.8 years of teaching experience, 7.5 in Houston.

As a group, these four teachers look like what we might imagine the typical inner-city school teacher to be. But dig deeper, and you'll find that until they came under the gun of EVAAS, their peers and supervisors viewed them as real Texas rangers battling great obstacles to educate children. And they were doing it well.

Teacher A was her school's 2010 Teacher of the Year. She had earned merit pay each year prior to her termination, and her principal recognized her success by giving her high marks across her evaluation scores, until the end. In her value-added scores, she showed a positive impact on her students' achievement just as many times as she showed a negative impact. Her "effectiveness" as measured by these scores was no different than what you might get through the flip of a coin, yet she was dismissed at least in part because of "a significant lack of student progress attributable to the educator," or "insufficient student academic growth reflected by value-added scores."

Teacher B showed two years of negative student test score growth and one year of positive growth in the three years leading up to her termination. Her value-added scores more or less matched her supervisor observation scores. This would seem to suggest that both the supervisor evaluation scores and the value-added scores measured her effectiveness validly. However, her last year was purportedly her best of the three that were used to determine her supposed ineffectiveness and, consequently, her termination. This trend challenges the position that she consistently demonstrated ineffective teaching.

Teacher C's value-added scores appear to clearly show that this teacher subracted value from her students' learning each year. Going by these scores, Teacher C seemingly did deserve to be terminated. However, she was also named Teacher of the Year in 2008 by her peers, she "exceeded expectations" on her principal evaluation scores across the same years, and she received merit pay every year. In addition, she taught some of the highest-need students in the district, some of whom were still in 6th grade even though they were almost old enough to drive.

Teacher D, like Teacher A, had value-added scores for three years that could have been produced by a coin flip. Then, within a single year, she surprisingly appeared to go from being a teacher with vacillating effectiveness to one of the worst teachers in the district. That same year, she had received a large influx of English language learners (ELLs) into her classroom. As with Teacher C, the characteristics of her students made it difficult for her to demonstrate added value, although the model claims to statistically factor out student demographics and their potential to bias scores (Wright, White, Sanders, & Rivers, 2010).

In response to their dismissals, Teachers A, B, and D began due process hearings, but they ultimately decided not to follow their hearings through to culmination. Instead, they decided to quit teaching in the Houston Independent School District or quit teaching altogether, saying, in true cowboy patter, that they were tired of the district's "bullshit."

Teacher C (the teacher who, according to her value-added output, demonstrated the poorest scores) took her case headlong and headstrong. Her hearing officer ruled that the high-needs students Teacher C taught most likely limited her capacity to add value, regardless of what the statisticians maintained. The hearing officer added that Teacher C did not have multiple years of consistent and statistically significant data in the subject areas she taught to warrant such a high-stakes decision.

Questions Data Can't Answer

A look at these four teachers, who faced the most significant value-added consequence possible, raises some interesting issues.

First, the evidence here and elsewhere shows that teachers who "add value" to their students' learning in one subject area may not appear as effective in another. Second, teachers who are successful at one grade level might not be successful in another, even if the grade levels are adjacent (for example, from grade 3 to 4). Ask yourself whether this makes sense. Then go a step further and hinge a teacher's job on the data available on these four teachers, and ask yourself what you would do if you were the new sheriff in town. What questions might you have?

For example, how could Teacher D suddenly become a bad teacher? Is it really possible to stop teaching well so quickly—or might this drop in the teacher's ratings have happened because the kind of students in her classroom changed? Weren't the statistical controls supposed to prevent such changes from affecting Teacher D's value-added rating? Perhaps the system does not work in the ways advertised. Perhaps a numerical equation cannot account for such contextual changes.

Consider Teacher C, who taught high-risk 6th grade students. Was it solely her fault that these students did not make the same level of progress as their peers, or were there other factors that might have contributed to their limited growth? Did these students care about the tests in the first place? (They were not held accountable, as their teacher was.) And what happened with Teacher A, whose value-added scores vacillated every year even as she was voted Teacher of the Year and awarded merit pay? Or Teacher B, who posted her highest value-added rating the year before she was labeled inept?

Unintended Consequences

By implementing its value-added model, Houston Independent School District has created a system that is deterring some teachers from teaching in classrooms in which they are most prepared to teach (Amrein-Beardsley & Collins, 2012). Houston teachers who have the opportunity to change teaching positions are becoming savvy about moving out of subject areas in which value-added measurement matters, moving to the grades in which it is easiest to show growth, or teaching students who are likely to test well.

As they have seen their jobs becoming increasingly jeopardized as a result of EVAAS-based decision making, Houston teachers say they have shied away from teaching English language learners, particularly in the 4th grade, when many ELLs are transitioned to English-only instruction. Teachers of gifted students express difficulties demonstrating growth as well, given the ceiling effects that the statisticians administering Houston's value-added model claim are irrelevant (A. Best, personal communication, January 21, 2012).

Although this study does not fully negate the possible benefits of value-added measures, it does call into question the purported benefits of using such measures for high-stakes decision making. The stated purpose of value-added models is to create an unbiased, objective measure of teacher effectiveness. These models do yield quantifiable, easy-to-understand numbers. But before we come out with guns blazing, ready to clean up the town and establish order as we hide behind the shield of our statistical formulas, we should remember that the use of value-added models is affecting more than just numbers. It is affecting real teachers with real lives—including the four earnest, dedicated, and previously respected teachers described here.

Even a superficial look at these four cases reveals that the value-added numbers are inconsistent with the other indicators of teacher quality. Such inconsistencies should raise a red flag to those who are a part of the system—including the SAS EVAAS developers themselves, Houston Independent School District administrators and school board members, and other district leaders who are in the process of creating or adopting their own value-added models.

Although these models might offer some "added value" to the ways we think about measurement of teacher performance, we must carefully consider their purpose and execution, along with their unintended effects, to avoid making more erroneous decisions that harm teachers and ultimately damage schools.

Authors' note: For a video describing the problems with value-added measures, go to http://education.asu/edu/evaas.

References

Amrein-Beardsley, A., & Collins, C. (2012). The SAS Education Value-Added Assessment System (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and unintended consequences. *Education Policy Analysis Archives, 20*.

SAS. (2012). *SAS®EVAAS®for K–12: Assess and predict student performance with precision and reliability*. Retrieved from www.sas.com/govedu/edu/k12/evaas/index.html

U.S. Department of Education. (2009). *Race to the Top program: Executive summary*. Washington, DC: Author. Retrieved from www2.ed.gov/programs/racetothetop/executive-summary.pdf

Wright, S. P., White, J. T., Sanders, W. L., & Rivers, J. C. (2010). *SAS®EVAAS®statistical models*. Retrieved from www.sas.com/resources/asset/SAS-EVAAS-Statistical-Models.pdf

Jessica Holloway-Libell is a doctoral student, Audrey Amrein-Beardsley is associate professor, and Clarin Collins is a doctoral candidate, Mary Lou Fulton Teachers College, Arizona State University, Phoenix.