The Journalism School
Knight Case Studies Initiative
Columbia University

CSJ-11-0040.0

Rating Teachers: a Job for Journalists?
The *Los Angeles Times* and "Value-Added" Analysis

By 2009, policymakers, educators and journalists who covered schools were increasingly frustrated with the nation's apparent inability to measure the effectiveness of its public school teachers. Standardized tests promoted nationally by the 2001 No Child Left Behind Act had begun to provide metrics of student performance. But who or what could determine whether an individual teacher was up to the job? To many who thought hard about public education and how to improve it, one possible solution was to track individual teachers according to their students' standardized test scores. This, they hoped, would make the educators accountable for their job performance.

At the *Los Angeles Times*, a reporting team decided to try just that. In November 2009, reporters Jason Felch and Jason Song and their editors obtained from the Los Angeles Unified School District (LAUSD) several years of elementary school standardized test scores in math and reading. The *Times* then hired an education economist to analyze the data and determine how individual teachers affected their students' test scores. By June 2010, the paper had generated a database of 6,000-plus elementary public school teachers that identified which teachers consistently raised students' standardized test scores, and which did not. Reporters were drafting a series of stories based on the data.

But simply having the information did not necessarily mean it should be published. The project had sparked many debates within the paper since its inception. There was discussion about whether it was appropriate for a news organization to rate teachers. Who were reporters to evaluate teaching? There were also questions about the methodology, so-called "value-added" analysis—a measurement approach that, while in use in several US school districts, had vocal critics. How could editors and reporters judge whether the results from this approach were trustworthy?

The thorniest question, however, was accountability: should the *Times* publish individual teachers' names? On one side stood those—mostly members of the team working on the story for nearly a year—who felt that tax dollars paid public teachers' salaries, and thus their work should

This case was written by Alice Irene Pifer for the Knight Case Studies Initiative, Graduate School of Journalism, Columbia University. (0711)

be subject to public scrutiny. They also argued that concrete rankings would allow parents to move their children from poor teachers to better ones, creating a virtuous circle that rewarded good teaching and perhaps encouraged less effective instructors to seek training or another career. On the other side stood some members of the *Times'* website staff, most of them newcomers to the project, who argued against naming teachers as an unwarranted invasion of privacy. To shine such a public spotlight on individuals was, they said, simply unfair.

In June 2010, Assistant Managing Editor David Lauter won approval from the top to move forward with the project. He was anxious to publish the series on teacher rankings before the school year started in September. But the objections gave him pause. Had the *Times* team been blinded in its editorial judgment by the triumph of designing what they considered a successful rating system? Had the team overlooked anything crucial? What would readers think? How would the paper handle any negative repercussions? Was it really a public service to name the teachers, or was the paper planning to publish names simply because it could?

Birth of the Project

The teacher ratings project grew out of a series that the *Los Angeles Times* ran in May 2009. Education reporter Song wrote several stories under the title "Failure Gets a Pass."[1] Song reported that LAUSD, due to union contracts, had difficulty firing the worst teachers—even those accused of molesting students. Instead, it transferred them out of the classroom but continued to pay them, which cost the LAUSD about $10 million a year.[2]

After the series ran, Education Editor Beth Shuster and Special Projects Editor Julie Marquis, who handled investigative stories, met to brainstorm about future education stories. They recognized that the problems laid out in "Failure Gets a Pass" applied only to a tiny percentage of LAUSD educators. What about less egregious, but still poor teachers? Perhaps the paper should next take a look at a persistent concern of parents—is my child's teacher helping him learn?

Teaming Up. In early June 2009, Marquis asked investigative reporter Felch if he wanted to work with Song on a story that probed which LAUSD teachers were effective. The two quickly discovered that the Los Angeles public school system had, in essence, no meaningful teacher rating system. What passed for evaluation was in-class observation once a year, or even only once every few years. Virtually every teacher was rated satisfactory. It was shocking, recalls Education Editor Shuster:

---

1 Jason Song, "Failure Gets a Pass," *Los Angeles Times*, May 2009. See: http://www.latimes.com/news/local/la-me-teachers-landing-html,0,1258194.htmlstory

2 New York City had a similar problem. See Steven Brill, "The Rubber Room," *New Yorker*, August 31, 2009. The city's worst public school teachers were assigned to so-called "rubber rooms," where they had no responsibilities but collected full pay while awaiting resolution of charges against them.

> Once we started really getting into this and learning that 98 percent of teachers are being rated as satisfactory, I mean, you know from working in organizations, any business, anywhere, 98 percent of the people are not satisfactory. They're just not.[3]

Felch and Song decided to research what was happening nationally. "We looked across the country, and it turned out that almost every state in the union was using a very similar approach, this kind of very rudimentary checklist to evaluate their teachers," says Felch.[4] As the two reporters took a deeper dive into the academic literature on teacher effectiveness, they found many experts who believed that the teacher was the single most important school-based factor in student success or failure—more important than class size or students' socioeconomic status. Moreover, they found that successful teachers did not necessarily cluster at high-performing schools but were scattered across the system. Similarly, even high-performing schools had struggling teachers.

One group of studies which reinforced the teacher-centric view was known as value-added analysis.

Value-Added Analysis

By the summer of 2009, several school districts were piloting the value-added model, including Dallas, Houston, New York City, Washington, DC and Chicago. Secretary of Education Arne Duncan was a supporter. When the Obama administration in July 2009 announced Race to the Top, a national initiative which allowed states to compete for federal funds to improve schools, it featured incentives to link teacher evaluations to student test scores.

Value-added metrics emerged from the world of economics. The model used complex algorithms to compute how much value a teacher added to his students' mastery of math and English as measured through standardized tests. The value-added model tracked individual teachers over the course of several years to determine whether their students' test scores consistently improved, declined or remained stagnant.

Support. Value-added supporters believed it could be a useful tool for school superintendents, principals and parents to hold teachers accountable. They argued that it controlled for socioeconomic differences among students because it rated teachers based on how much their students improved. Thus, the teacher was judged not on standard grade level expectations for students, but on their progress. Teachers in low-income areas, for example, with

---

3 Author's interview with Beth Shuster on March 29, 2011, in Los Angeles. All further quotes from Shuster, unless otherwise attributed, are from this interview.

4 Author's interview with Jason Felch on March 29, 2011, in Los Angeles. All further quotes from Felch, unless otherwise attributed, are from this interview.

fewer students performing at grade level, could still receive an effective rating so long as the majority of students made significant progress.

Criticism. The method's critics, however, charged that its calculations were misleading and often wrong. For one thing, researchers in each jurisdiction were free to decide which variables to include or exclude—for example, student family income, parent educational level, race, or proficiency in English. Thus, the same set of data could generate different results based on the variables selected. This also meant that results could not be compared across states or districts because the methodology was not consistent. The model also failed to factor in important information such as whether a class was team-taught, or whether a student or teacher had been absent for prolonged periods. Finally, it depended for its findings on standardized tests, whose own validity had been the subject of intense debate for decades.

One issue that both supporters and critics agreed on was that value-added should be only one component in a teacher's evaluation. They differed, however, on the weight it should be given, with estimates ranging from as high as 50 percent to a low of 3 percent.

At the *LA Times*, reporters and editors knew that value-added analysis had its limitations. But the more Felch and Song researched it, the more they came to believe that it was likely the best method available for assessing a teacher's abilities. As Felch says: "This was the key that these researchers were using to kind of unlock this world, where we suddenly were able to see dynamics that were going on that had been blurred before." They decided to see whether the value-added approach could possibly work for teachers in Los Angeles public schools.

Do it here? To do a valid value-added analysis, researchers required several years of continuous student test scores linked to their teachers. The LAUSD had been collecting data from the California Standards Tests (which it adopted in 2002) for seven years. Felch and Song proposed to their editors, Marquis and Shuster, that the *Times* try to obtain the LAUSD data. If successful, the *Times* could hire an expert in value-added analysis to rate the LAUSD teachers. The paper could then post the results on its website, along with a series of articles putting the data in context. To increase the chance of influencing real policy change, the paper might even name the teachers.

Marquis and Shuster thought the idea had potential, but believed it unlikely LAUSD would release the data. AME Lauter agreed that the prospects for LAUSD cooperation were slim. But he loved the idea, and felt that if they could get the data, the *Times* would be in a position to provide a valuable public service.

Lauter also worried about the cost, especially the expense of an outside consultant. While *Times* management wanted to support ambitious journalism, 2009 had been a particularly bad year financially for newspapers. Nonetheless, Lauter advised the reporters to push ahead. In the meantime, he applied for funding to the Hechinger Institute on Education and the Media, an arm

of Teachers College at Columbia University that supported major journalism projects focused on education.[5]

Seeking the Data

The editors first approached individual LAUSD district officials for the student test results, but none would release the data. So in late July 2009, Felch and Song went to the top: Superintendent Ramon C. Cortines. When they met, Felch and Song had armed themselves with numerous arguments to convince the superintendent to give them the raw scores. Much to their surprise, Cortines said yes. Yet this did not mean the *Times* got the data right away. Special Projects Editor Marquis wasn't surprised. She says:

> There's a difference between getting the superintendent to say, sure, you can have it, and getting the school district lawyers and all the bureaucrats underneath him to release it to you, because there are federal laws protecting student privacy, and there are sometimes different agendas at different levels of an organization.[6]

On October 5, the *Times* filed a formal request for the data using the California Public Records Act. Meanwhile, there were other issues on the table.

*Appropriate Role***?** From the moment the Felch-Song proposal began to work its way up the ranks at the *Times,* there were ongoing discussions about whether it was appropriate for the *Times* or any news organization to rate teachers. After all, wasn't that the responsibility of the LAUSD? But Song and Felch learned that, as long ago as 2006, an internal LAUSD report had recommended using value-added to evaluate teachers. The school district, however, had ignored the recommendation for fear of complicating ongoing union contract negotiations.

The newspaper was under no such constraints. Parents, *LA Times* staff believed, had a right to know about the effectiveness of those teaching their children. Informing the public, including parents, was the mission of a newspaper. Many felt the *Times* would be performing a service that, for political reasons, neither the LAUSD nor the teachers' union had taken on. AME Lauter, for one, found the project worthwhile. He says:

> It seemed to me that although the task would be complicated, that if we could do it in a solid, accurate, meaningful way, that this could be a really important step towards transparency. That's something that is really at the heart of a news organization's role.

---

[5] The Hechinger Institute in August 2010 awarded the *Times* a $15,000 grant that the paper used to help defray the cost of the consultant.

[6] Author's interview with Julie Marquis on March 29, 2011, in Los Angeles. All further quotes from Marquis, unless otherwise attributed, are from this interview

Finally, in November 2009, the *Times* received a first round of data from the school district, and then a cascade of material. The *Times* had agreed to various LAUSD conditions, such as protecting student privacy. Rating teachers was no longer a theoretical possibility; it was real. Reporter Felch was excited: "Suddenly we had this massive data… and we knew the power of what could be done with it. So then we set about doing it."

Earlier in the fall, Felch had researched consultants who could conduct the value-added analysis should the *Times* get the data. He recommended Richard Buddin, a respected education economist from the RAND Corporation. Buddin was an expert on teacher performance, teacher evaluation, and value-added analysis. In November, with the data in-house, it was time to bring Buddin on board to get the project underway.

In December 2009, AME Lauter gave Editor Russ Stanton his first full briefing on the project. Stanton was enthusiastic, convinced that this could be an important contribution to watchdog journalism in Los Angeles. He encouraged Lauter to proceed.

Creating the Ratings

As Buddin began to work with the LAUSD information, he first created a pool of raw data: the standardized test results for students in grades two through five. The data covered some 603,500 elementary students taught by about 18,000 teachers in 520 schools. Buddin could not use it all, however—he was interested in the student test results of English and math teachers teaching 3rd-5th grade (testing started in second grade, and Buddin needed at least two consecutive years of test scores to conduct a value-added analysis). That gave him a pool of some 6,000 teachers.

The data, he found, included information on students' gender, age, poverty level, number of years in the LAUSD, and whether a student was a non-native English speaker (the *Times* requested additional demographic data on race and ethnicity, but the LAUSD refused due to privacy laws). The test results could be sorted by teacher as well as by school, type of school (standard or charter), and grade level.

Buddin used complex mathematical formulas and regression analysis to try to determine what effect individual teachers had on their students' learning over time. He posed three questions he hoped the data could help answer: how much did teacher quality vary from school to school and from teacher to teacher; what qualifications or background influenced teachers' success in the classroom; and how did traditional measures compare with value-added measures of teacher and school effectiveness?[7] He hoped in particular that his analysis would help document whether standard teacher credentials—advanced degrees, special training, or years of experience—correlated with the achievement of their students in the classroom.

[7] For more detail, see: http://www.latimes.com/media/acrobat/2010-08/55538493.pdf

As insurance, Data Analysis Editor Doug Smith proposed (and Lauter approved) that he create a parallel system to double check Buddin's results. While Smith's set-up was not as sophisticated as Buddin's, he had the capability to analyze the LAUSD data. Smith amplifies:

> We did a simple gain-score analysis. That is, we ranked the students in each grade level into percentiles, calculated each student's change in percentile from year to year, and summed those differences for each teacher's students. We then ranked the teachers into quintiles based on their average student gain scores. Then we repeated the process for schools.[8]

Smith and Buddin continually compared the two sets of results.

*Crisis of Confidence***.** By early 2010, Buddin had some preliminary findings for particular teachers. To test these, reporter Felch in late January spent two days at the Carpenter Avenue Elementary School. He went specifically to check on one teacher who, according to Buddin's calculations, had a very low score. Felch wanted to see if Buddin's theoretical results matched with the teacher's actual performance in the classroom. During his visit, Felch saw engaged students and what looked to him like a good teacher. He also spoke with the principal and other teachers; there was no sense that anything was amiss with the teacher in question. Felch recalls:

> I came away with grave concerns about the quality of the data. I came back, and I said, "You know, guys, this ain't it. If this is what the data's telling us, I don't think it's really that valuable."

Felch sat down with Buddin and Smith to try and figure out what was wrong. Buddin pulled apart his complex statistical analysis, and quickly found a major error in arithmetic. As it turned out, fully one-third of the teachers had received erroneous scores, including the one Felch had visited. While a significant mistake, it was easy to fix.

Other fixes followed. For example, Smith and Buddin had made different decisions about which teachers and students to include or exclude. Buddin, coming from an academic background, typically worked with very large databases and was able—without distorting the results—to discard any data deemed possibly unreliable. So he had left out test results from charter schools. But Smith wanted as large a pool as possible in order to improve the accuracy of the analysis. While the *Times* had plenty of district-wide data, the amount of data linked to any individual teacher was small; just a few missing students could affect a teacher's rating significantly. So they agreed to include charter schools. In general, Smith restricted discards to demonstrably inaccurate data—for example, students who had two math scores for the same year, or a teacher who was listed at two schools.

[8] Excerpt from Doug Smith email to author, June 14, 2011.

There were other adjustments. Buddin and Felch (before Smith joined the project) had originally agreed to a standard for inclusion in the analysis: a teacher had to have taught at least 60 students, and those students had to have had at least two consecutive years of standardized test scores in math and English. For some reason (Smith suspects himself of carelessly saying "greater than 60" rather than "greater than or equal to 60"), Buddin included only teachers who had taught *more than* 60 students. That small difference had eliminated a surprising number of eligible teachers. When Buddin shifted to the "at least 60" standard, it meant more teachers in the database, and changed the rating for many.

In another example, in comparing notes on teachers for whom Buddin and Smith had different scores, they discovered that each had processed test results for English Language Learner (ELL) students in a different way. Buddin had boosted the ELL students' test scores in order to compensate for classes with numerous ELL students. Without the boost, it would have been unfair to compare teachers with no ELL students to those who had a significant number (the boost raised the teachers' scores). Smith realized that Buddin was correct on this point and changed his own methodology. Smith also discovered that he had mistakenly included in the pool some teachers who had transferred to middle school, beyond the scope of the study, and removed them.

With these major problems resolved, the project looked like it had a good chance of working. So on February 14, 2010, Lauter and his team set a budget which included additional compensation for Buddin. Importantly, on the same day Lauter secured Editor Stanton's agreement that if the value-added analysis did not generate solid results, the *Times* would not feel obligated to run a story. That meant considerable risk: the paper might invest substantial resources and nevertheless hit a wall. But Stanton was willing to take the risk. They would publish only if the product was credible.

More Reality Checks. By late March, Buddin completed the teacher ratings. But how to test them in the real world? Editors Marquis and Shuster came up with a solid idea, says AME Lauter:

> We started talking about how are we going to test this? What we early on decided was, if we could find teachers who were clearly disparate in their ratings [yet] who were teaching very similar kids at similar schools, that that would illustrate how the method works in a way that's very intuitively understandable for people… So we started looking for those sort of matched pairs of teachers.

Once they identified the pairs, reporters Felch and Song during April and May 2010 visited schools to observe and conduct interviews. They observed more than 50 elementary school teachers in over a dozen schools. They found that Buddin's ratings matched with their reporting on the ground. Song, for one, was reassured. He had never been fully confident that value-added analysis would produce more than theoretical results: "I wasn't sure whether it was going to work,

whether the value-added system even had any kind of relevance to reality."[9] He was glad to see it did.

Data Analysis Editor Smith, too, was satisfied that the value-added model was generating reliable results. But the ratings themselves disturbed him. Most teachers, about 80 percent, received average ratings; another 10 percent were rated highly effective. But fully 10 percent were rated highly ineffective. Smith and Buddin found that students of these teachers dropped seven to 15 percentage points in their test scores. Smith says:

> I got sick in my stomach looking at what happened to the students who had the worst teachers. The differences were assaultive. I mean, it wasn't on a gray scale. These students that got the worst teachers were diving.[10]

With these results, AME Lauter and his group of editors and reporters believed more strongly than ever that it was their responsibility as journalists to get this information to the public. Still, the question remained: should they name individual teachers?

Name the Teachers?

For months, there had been intermittent discussions about whether to name the teachers. The team did consider other options. For instance, the *Times* could just report the number of teachers in various categories at individual schools—highly effective, average, highly ineffective—without naming names. In effect, rate the schools. But Special Projects Editor Marquis, for one, saw no point in that. She elaborates:

> Then what are you telling people? OK, people, 20 percent of the teachers in this district are really bad at raising students' test scores, and we know who they are. But we're not going to tell you, because you might misunderstand it? I mean, that's what I don't get.

Reporter Song started out more wary but came to agree with Marquis that the *Times* couldn't publish a story saying it had identified good and bad teachers, and then not provide readers with the underlying data. As a journalist, he believed that, in general, reliable information about public employees should be disclosed. Education Editor Shuster also favored naming the teachers. As a mother, she knew that parents usually found out about teachers through "gossip in the parking lot." In a perfect world, she felt that LAUSD should have provided parents with solid teacher evaluations instead of leaving them to rely on rumors. Now the *Times* could do what the

[9] Author's interview with Jason Song on March 28, 2011, in Los Angeles. All further quotes from Song, unless otherwise attributed, are from this interview.

[10] Author's interview with Doug Smith on March 30, 2011, in Los Angeles. All further quotes from Smith, unless otherwise attributed, are from this interview

school district had not. AME Lauter initially was more skeptical, although in time he came to see naming teachers as a public service. He explains:

> My grandparents were public school teachers in New York. My parents are college professors. So I'd kind of grown up around teachers. My initial reaction was, well, I don't know whether I'm comfortable doing this.

In fact, many involved with the project had teachers in their families. Editor Stanton's oldest daughter was a public school teacher, and his mother had been one as well. Reporter Felch had been a teacher himself before he became a journalist. Felch fully understood that some teachers would be embarrassed, and their reputations damaged, if the *Times* rated them low. But he believed that the greater good to the community outweighed this concern.

A lingering skeptic was Data Analysis Editor Smith. Smith did a lot of soul searching about posting teachers' names. He was a 40-year veteran at the *Times* and had become the dean of computer-assisted reporting. He knew firsthand the power and the limits of data. All data, he knew, were "dirty" (misleading or incorrect) to one extent or another and, if not handled properly, flawed data could give unreliable results. But after the lengthy period of processing the data, and matching value-added results against classroom reality, Smith came around. He favored publishing the names.

But it was not Smith's call. AME Lauter decided it was time to meet again with Editor Stanton.

Green Light?

On June 16, Stanton hosted an hour-long meeting in his office with AME Lauter, Data Analysis Editor Smith, Special Projects Editor Marquis, and Education Editor Shuster. The purpose was to decide whether the paper would run the story or not. So first, the team updated Stanton and made the case for going ahead.

But Smith wanted to make sure there were no unexpected surprises, either for the *Times* or for its editor. So, with Lauter's approval, he presented the case *against* doing the project, laying out all the potential pitfalls. The team fully expected the series to cause a lot of controversy. Smith listed the drawbacks. For instance, he said it was inevitable that with a database this large, some teachers would be rated incorrectly—some to their advantage and others to their disadvantage. He noted that while he and Buddin had done everything possible to minimize errors, it would be impossible to eliminate them all.

Stanton grilled the staff about the methodology of the value-added analysis, as well as the final results. After some back and forth, Stanton declared himself satisfied. He judged both the methodology and the teacher ratings valid and defensible.

But what about the most sensitive question? When Lauter asked Stanton whether they should publish the teachers' names, there was a very long pause. Then Stanton, a father himself, gave his decision—name the teachers. "I went into parent mode. If I was a parent, I would be mad at the *LA Times* if they did all this stuff and said here's all these things, but we can't tell you who they are because we want to respect their privacy or what have you," comments Stanton.[11] The project was a go.

Internal Resistance

Even with Stanton's approval, there remained much to do before publication. Building the *Times'* website capacity was key. While Felch and Song's stories would appear both in the print and online versions of the *Times*, the website was the only place where the enormous database listing the 6,000-plus teachers and their ratings could be posted. It was time to involve the data team. On June 18, Lauter explained the project to Online Deputy Editor Megan Garvey, who would oversee the effort. She was dismayed. She says:

> He tells me, and I immediately have almost a visceral reaction, honestly, against it. I'm thinking, well, are we validating test scores as the only way to measure whether a teacher is good?[12]

Garvey worried especially whether it would be fair to name the teachers. She had worked at the *Times* for 12 years and knew and trusted many of people on the project team. Yet she didn't like the sound of this undertaking. Neither did Web Producer Ken Schwencke, whose assignment was to design the online presentation of the teachers' names and ratings. He said that if he didn't believe in the project, he would refuse to create the database. Schwencke explains:

> I was more worried about the accuracy of the methodology than anything. At that point, I had never heard about value-added methodology. I think there is probably initial discomfort whenever you talk about disclosing job performance, basically about people. But my main thing was, I didn't understand the methodology.[13]

Editor Stanton, AME Lauter, other top editors, and reporters Felch and Song had already considered these issues during the year-long build-up, and were ready to move ahead with the story. It was now mid-June, and Lauter wanted it published before September. He had not

---

11 Author's interview with Russ Stanton on March 31, 2011, in Los Angeles. All further quotes from Stanton, unless otherwise attributed, are from this interview

12 Author's interview with Megan Garvey on March 30, 2011, in Los Angeles. All further quotes from Garvey, unless otherwise attributed, are from this interview

13 Author's interview with Ken Schwencke on March 31, 2011, in Los Angeles. All further quotes from Schwencke, unless otherwise attributed, are from this interview.

expected Garvey and Schwencke's reactions. He had no doubt that he could get the work done in time, either by ordering his team to do it or by hiring outside programmers.

But their questions gave him pause. Were they a harbinger of how the public would react? What might the fallout be from publication, and how should the editorial leadership prepare itself? Above all, had the *LA Times* leadership made the right calls, or should it reevaluate?

This case was written by Alice Irene Pifer for the Knight Case Studies Initiative, Graduate School of Journalism, Columbia University. The faculty sponsor was Professor LynNell Hancock. Funding was provided by the John S. and James L. Knight Foundation. (0811)