WHAT DID THE TEACHERS THINK? TEACHERS' RESPONSES TO THE USE OF VALUE-ADDED MODELING AS A TOOL FOR EVALUATING TEACHER EFFECTIVENESS

Linda Lee
California State University, Los Angeles

ABSTRACT

The policy discourse on improving student achievement has shifted from student outcomes to focusing on evaluating teacher effectiveness using standardized test scores. A major urban newspaper released a public database that ranked teachers' effectiveness using Value-Added Modeling. Teachers, whom are generally marginalized, were given the opportunity to respond to their rankings. This research examines a subset of those teachers' perceptions about the use of standardized test scores in determining teacher effectiveness. It is important for policy makers to hear from those whom are the implementation level of such major policy shifts in education reform.

Keywords: Teacher Effectiveness; Teacher Attitudes; Teacher Response; Teacher Evaluation; Evaluation Methods; Value-Added Models; Accountability; Educational Policy; Elementary Education

Introduction

In August 2010, a major urban newspaper, the Los Angeles Times (*L.A. Times*), published a study on teacher effectiveness using a statistical method, Value-Added Modeling (Buddin, 2010). The results of the study were published in an online database, which showed individual rankings of teacher effectiveness, based on the teacher's students' progress on standardized test scores in English and math. The "value" a teacher adds or subtracts is based on the difference between a student's expected growth and actual performance on the tests. The database included about 6000 Los Angeles Unified School District teachers that taught at least 60 students in the third, fourth and fifth grades, during the 2003 to 2009 school years. The newspaper's statement on the purpose of publishing the information was "...it bears on the performance of public employees who provide an important service, and in the belief that parents and the public have a right to judge it for themselves" (Felch, et al., 2010).

The public release caused a stir, because, for the first time, the public was able to see quantifiable differences amongst teachers. In tandem with the release of rankings, the newspaper

Linda Lee is a doctoral candidate in the EdD Program in Educational Leadership at California State University, Los Angeles and an administrator at an urban elementary charter school. Ms. Lee can be reached at CSU Los Angeles, Division of Applied and Advanced Studies in Education, 5151 State University Drive, Los Angeles, CA 90032. E-mail: llee18@calstatela.edu.

gave teachers the opportunity to respond to the rankings and use of test scores in evaluating teacher effectiveness. In doing so, the L.A. Times provided the public with a rare opportunity to hear from the teachers, whom often when decisions on educational policy are made, are left out of the conversation. This is powerful in the sense that by ―searching the margins…one finds the great potential of people expressing counter narratives and alternative proposals for policy" (Marshall & Gerstl-Pepin, p. 152). In the responses posted, teacher gave opinions, arguments, and suggestions about the use of Value-added Modeling. The purpose of this study is to analyze these responses, so that we can better understand some of the challenges and nuances of trying to measure a process as dynamic as teaching and learning. Understanding the teachers, who are the negotiators of the transactions between teaching and learning, is essential to illustrate some of the challenges the nation faces as it moves to evaluating and rewarding effective teachers, and, ultimately, the implications for producing educated citizens.

Unfortunately, effective evaluation of teachers has been an elusive task, where we have lacked the ability to discern effective and ineffective teachers. Weisberg, Sexton, Mulhern, & Keeling's (2009) study of twelve districts in four states showed that, in districts with binary evaluation ratings (satisfactory/unsatisfactory), more than 99 percent of teachers received a satisfactory rating. In districts with a broader range of ratings, 94 percent of teachers received one of the top two ratings and less than one percent received an unsatisfactory rating. A study on statewide policies on teacher evaluation in the mid-west region (Brandt, Thomas, & Burke, 2008) found that most states provided guidance to districts on evaluating their teachers, which included criteria ranging from who is responsible, to frequency of evaluation. However, the criteria were general to the status of the teacher, rather than teaching and learning. Similarly, the No Child Left Behind Act provided the requirement of having Highly Qualified teachers, but the qualification only went so far as tracking credential status. Meeting the definition of Highly Qualified neither predicted nor ensured that a teacher would be successful at increasing student learning.

In addition to having ineffective evaluation tools, efforts to increase student learning have been challenging. According to the National Center for Educational Statistics (Rooney et al., 2006), since the early 1990s, the achievement gaps between White and Black, and White and Hispanic, have shown little measurable change. The inability to close these gaps has resulted in looking beyond student achievement on standardized tests and is now sharply focused on teachers. The basic framework of logic, which is driving much of the nation's current efforts in closing the achievement gap, is the notion that if you have good teachers, you will have good student achievement. Or, one can inversely infer: bad teachers are preventing our students from achieving. This notion of having teachers with different levels of effectiveness has become a major focal point in federal government's plan to ―fix" the problem of low student achievement. The Blueprint for Reform (US Department of Education, 2010) ties teacher effectiveness with student test scores:

―We will elevate the teaching profession to focus on recognizing, encouraging, and rewarding excellence. We are calling on states and districts to develop and implement systems of teacher and principal evaluation and support, and to identify effective and highly effective teachers and principals on the basis of student growth and other factors." (p. 4). This has led to a drive to find a way to measure teacher effectiveness using standardized test scores as the tool.

Value-Added Modeling

One statistical method that policymakers see as a tool for teacher evaluations is Value-added Modeling (McCaffrey, Lockwood, Koretz, & Hamilton, 2003), a statistical method that calculates individual student growth by comparing his/her previous year's test score to his/her current year's score, and comparing that growth in relation to other students in that grade level. Policymakers around the nation are embracing the idea of using a value-added measurement tool because it seems to provide an objective measure in evaluating teacher effectiveness. However, researchers have cautioned the use of Value-Added Models (VAM) due to limitations and unsolved problems. For instance, Schochet & Chiang (2010) found that more than 90 percent of the variation in student gain scores is due to the variation in *student-level* factors, and strongly suggests that policymakers carefully consider system error rates in designing and implementing teacher performance measurement systems that are based on value-added models. Another factor, is the issue of missing data (van de Grift, 2009), where the results are only valid for the detection of schools with the highest raw scores and the highest learning gains. In addition, Papay (2011) found that the different tests did not rank individual teachers consistently. Because of these and other limitations, Baker et. al. (2010) argue that VAM should only be one component, and a comprehensive evaluation should be standards-based and include evaluation by supervisors and peers. Thus far, the discourse on determining teacher effectiveness with the use of VAM has mainly been at the policy and research levels. We need to solicit teacher perspectives to understand the subtleties involved with evaluating teaching and student learning. However, there are few conduits of influence where teachers can have their opinions heard. Often times, their viewpoints are mediated through others (e.g. unions, administrators, associations) or not surfaced at all for the knowledge of the general public. Including teachers in the discourse is essential, as it can provide valuable information from those that are directly charged with increasing student achievement, information that would normally be missed when making policy decisions. Hence, this study will analyze the teachers' responses to the use of VAM in determining teacher effectiveness.

Research Question

What are the perceptions of teachers who are working in a large urban school district concerning the use of VAM in evaluating their effectiveness?

Sub questions: Do teachers differ in their opinions based upon their individual rankings? Is there a relationship between Overall Ranking and Years of Teaching Included?

Methodology

This is a mixed methods study that utilizes non-participant observation strategies through an unobtrusive research design due to the fact that the data set is publicly posted on the Internet. As of December 2010, 293 teachers posted responses. Only teachers who were part of the released rankings were allowed to post a response. Information collected from the database included: the submitting teacher's name, the time and date of the submission, teacher's VAM Overall Ranking, VAM ranking in English, VAM ranking in Math, number of years included in the ranking, the school they were employed at during the most recent standardized test administration, the schools where they were previously employed, and the teacher's response.

Each response was analyzed to determine whether the teacher was generally positive/agreed with the use of VAM, negative/disagreed, or neutral/mixed. For quantitative analysis, a frequency count determined the number of respondents at each of the five levels of rankings, ranging from least effective to most effective. Cross-tabulation was used to categorize the type of comment (Positive/Agree, Negative/Disagree, Neutral/Mixed) within each level of ranking. In addition, a correlation analysis examined teacher rankings in relation to the number of years teaching included in the study. Qualitatively, conventional content analysis (Hsieh & Shannon, 2005) was used to allow for categories to emerge from the data. As the responses were being read through, open coding was used to select content by marking key words, phrases, sentences and paraphrases of the responses. Units of code, ranging from single words to sentences, were gathered and then sorted into related categories. Several common categories were determined from the patterns of the units (e.g. arguments, opinions, outcomes, alternatives, etc.). These were then grouped into three main categories to determine common elements in the responses: knowledge, attitudes, and beliefs.

Findings

Quantitatively, frequency counts of each type of respondent (i.e. least effective, less, average, more, most effective) demonstrated a range of 17.4% - 22.8%, which is approximate to the quintile breakdown used in VAM. Hence, there was a fair balance of responses from teachers at each of the five ranking levels. Upon analyzing the nature of the responses, it was found that the majority of the responses (221 of 293) were categorized as Negative/Disagree (see table 1). The level that had the most categorized as Positive/Agree was the "Most Effective" level, where many responses indicated that the teachers were appreciative of having recognition of their efforts. Notably, although this level had the most positive/agree responses, the majority of the responses were negative/disagree towards the use of VAM.

Table 1

Cross-tabulation of Overall Rank and Type of Comment

		Type of Comment			
		Negative/Disagree	Neutral/Mixed	Positive/Agree	Total
Overall Rank	Least	45	2	3	51
	Less	50	4	7	61
	Average	48	9	8	65
	More	42	6	4	52
	Most	35	4	25	64
Total		221	25	47	29

An evaluation was made of the relationship between Overall Rank and years of teaching within the 6-year window using Pearson's correlation. The analysis showed that the results were not statistically significant, $r = .100$, $p > .05$. Therefore, no relationship between the ranking of the teacher and the years of teaching that were included could be determined (see table 2).

Table 2

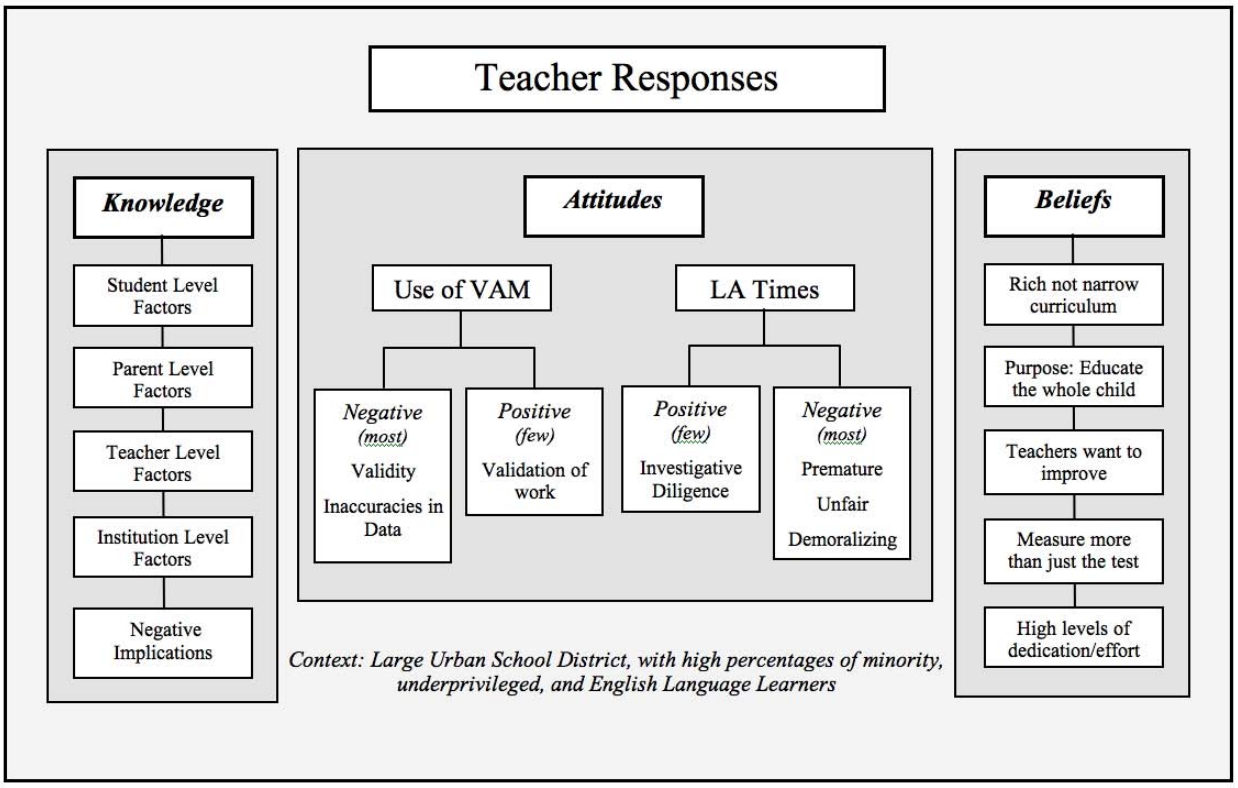
Correlation Analysis of Overall Rank and Years Included

		Overall Rank	Years Included
Overall Rank	Pearson Correlation	1	.100
	Sig. (2-tailed)		.088
	N	293	293
Years Included	Pearson Correlation	.100	1
	Sig. (2-tailed)	.088	
	N	293	293

In using conventional content analysis, initially, over 850 codes emerged through open coding. From the codes, more than 300 patterns of text were identified. These patterns were then categorized into themes. Major themes were then classified into three categories: the knowledge, attitudes, and beliefs teachers had regarding the use of VAM for evaluation of effectiveness (see figure 1).

Figure 1

Teachers' Knowledge, Attitudes, and Beliefs



Overwhelmingly, teacher attitudes towards the use of VAM was negative due to what they perceived as a disconnect in defining the education of the whole child with a test score in English and math. In particular, strong affective terminology was most used with regard to the public release of teacher names and rankings (e.g. demoralizing, resentment, public stoning, offensive, irreversible.) criticizing how the information was disseminated, and the lack of privacy for teachers. Many teachers were angered and felt that the newspaper was premature, irresponsible, and unfair. The responses also demonstrated that teachers had knowledge that validated many of the issues that already exist in the literature, such as the impact of student-level factors (e.g. special education students, students with little room to improve, English Language Learners), parent-level factors (e.g. education level, support at home), teacher-level factors (e.g. team teaching, previous teacher effects, being on leave for part of the year, teaching to the test), and institution-level factors (e.g. type of curriculum, leadership, lack of random assignment of students). Implications that were raised included: increased competition amongst teachers; under-performing children being ~~"u~~nwanted"; ~~"b~~randing" teachers; narrowing of the curriculum; cheating as a means to ~~"g~~ame" the system; and parental competition for those labeled as most effective teachers. Concepts introduced by teachers included: lack of recognition of their dedication and efforts; lack of resources to properly teach; influences of school culture; influence of teacher seniority on selection of classes; influence of school initiatives and programs; interference of district and union policies; year-round vs. traditional calendars; importance of administrator competence; degradation of the level of collaboration found in professional learning communities; restrictive curriculum; and influence of lack of student motivation for doing well on the test. Teachers' beliefs surfaced issues about necessity of having a rich curriculum to develop a whole child, the purpose of education being the educating of an individual not a test score, the turning of education into a business model, and that teachers want to improve in their practice. Responses indicated that teachers welcomed a process for evaluation to improve practice, but it should be done privately, and that VAM should not be the sole tool for evaluation. They suggested including other measures such as classroom observations, parent feedback, student feedback, and portfolios.

Further investigation is warranted to understand what metrics teachers would apply to the things they deem important in the education of a child. Also, some teachers indicated the need to remove ineffective teachers, but what was lacking in the responses was how to identify ineffective teachers. Further study is needed in order to understand what criteria teachers would use to determine ineffectiveness, and whether those criteria would be similar to ones used to identify effectiveness. In addition, there is little reference in the literature to the issue of the social learning environment. The process of learning is not isolated to the relationship between the teacher and an individual student. Rather, learning is also constructed upon interaction with peers, and is a dynamic process that is also dependent upon inter-relationships and interactions within and outside the classroom. Because these teachers work in an urban district that serves high percentages of minority, underprivileged, and English Language Learners, further exploration is needed how effectiveness can be measured when the challenges are compounded.

In conclusion, this study found that teachers identified many factors (e.g. institutional, teacher, parent and student level), which are outside of a teacher's control, that influence who and how they teach. Hence, the use of standardized test scores is not a valid measurement of teacher effectiveness. Most significantly, they argue for an evaluation that addresses the development of the whole child by fostering critical thinking, love of learning, and respectful

citizenship, through a rich and diverse curriculum. An implicit assumption that can be made from their responses is that what VAM measures is not aligned to what teachers see as the purpose of education. This misalignment stems reform efforts in which there has been a substantial change in our purpose of education, where we have moved from the development of the individual as a basis for a democratic society, to the development of individuals as a currency for economic competitiveness. This misalignment is noteworthy for all of us, because society's definition of the purpose of education ultimately affects the type of educated citizen that is produced, and how that education is measured.

References

Baker, E. L, Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*. (EPI Briefing Paper #278). Washington, DC: Economic Policy Institute. Retrieved from http://www.epi.org/publications/entry/bp278

Brandt, C., Thomas, J., & Burke, M. (2008). *State Policies on Teacher Evaluation Practices in the Midwest Region. REL Technical Brief. REL 2008-No. 004*: Washington, DC: Institute for Educational Sciences.

Buddin, R. (2010). *How Effective Are Los Angeles Elementary Teachers and Schools?* Retrieved from http://documents.latimes.com/buddin-white-paper-20100908/

Felch, J., Ferrell, S., Garvey, M., Lauder, T. S., Lauter, D., Marquis, J., Pesce, A., Poindexter, S., Schwencke, K., Shuster, B., Song, J., & Smith, D. (2010). Los Angeles Teacher Ratings. *Los Angeles Times.* Available online from http://projects.latimes.com/value-added/

Hsieh, H. F. & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research 15*(9), 1277-1288

Marshall, C. & Gerstl-Pepin C. (2005). *Re-Framing Educational Politics for Social Justice*. New York: Pearson Education, Inc.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., Hamilton, L. S. (2003). *Evaluating Value-Added Models for Teacher Accountability*. Santa Monica, CA: Rand.

Papay, J. P. (2010). Different test, different answers. The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193.

Rooney, P., Hussar, W., Planty, M., Choy, S., Hampden-Thompson, G., & Provasnik, S. (2006). *The Condition of Education*, 2006. NCES 2006-071. Washington, DC: National Center for Education Statistics.

Schochet, P. Z. & Chiang, H. S., National Center for Education, E., & Regional, A. (2010). *Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains. NCEE 2010-4004*. Washington, DC: National Center for Education Evaluation and Regional Assistance.

US Department of Education. (2010). *A Blueprint for Reform: The Reauthorization of the Elementary and Secondary Education Act*. Washington, DC: Author.

van de Grift, W. (2009). Reliability and validity in measuring the value added of schools. *School Effectiveness and School Improvement, 20*(2), 269-285.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness*. Brooklyn, NY: The New Teacher Project. Retrieved from http://widgeteffect.org