Teacher Evaluations: Problems and Possibilities

Douglas F. Warring
College of Education, Leadership & Counseling
University of St. Thomas, Minneapolis, MN 55403 USA

Abstract

Across the United States and internationally states, districts, and schools are in the process of developing teacher evaluation systems. The measures being used to determine the future of P-12 teachers often do not utilize appropriate multiple measures and factors that need to be taken into account in the evaluation of teachers. While the evaluations are designed to improve teacher practice they often do not take into account the fact that in a single environment, learners and teachers themselves vary in beliefs, attitudes, perceptions, self-efficacy, motivation, learning styles, cultural influences, and demographics or social identities (e.g., sex, sexual orientation, ethnicity, ability/disability, socio-economic status, religion/spirituality, etc.). Typically the many levels of diversity just noted are typically not considered in calculating the teacher evaluations. This paper advocates use of multiple measures in evaluation of teachers and suggests key issues that need to be considered in teacher evaluations.

1. Introduction

In recent years schools have been required to submit standardized assessment data based on the assumption that assessment data can provide credible information on progress and judge the quality of the teachers. Most people understand that one-time assessments are not a fair way to assess learning or to evaluate teachers. Evaluations systems that are well designed take into account many significant variables offer tools for growth and effective teaching.

Teachers are the most influential school-based factor on student achievement [1] [2] [3]. Student learning growth should be measured with sophisticated statistical models and student learning is a desired outcome. Testing systems should be organized to support teachers in their efforts to improved instruction. An over reliance on standardized tests over-emphasize testing and often do not take into account the important work of teachers in teaching and student learning.

Although studies have shown that some teachers are more effective than others at helping their students achieve at high levels, most indicators of teacher quality (e.g., credentials, characteristics, and observable practices) are generally poor predictors of student learning growth [4] [5] [6]. Teachers' scores on observation instruments have not been highly correlated with student learning growth [7]. However, it is not surprising that correlations are weak when the factors to be measured with observations are not well specified, or when raters are poorly trained or inadequately monitored for scoring consistency after training.

Improving teacher effectiveness is a very complex issue. The ability to identify teachers who are performing well as those who are not performing well is crucial in providing the appropriate staff development. Unfortunately teacher evaluations have varied widely and often differ significantly in their breadth and scope. P-12 education, in addition to examining teacher impact on student achievement, should also be concerned with many issues paramount to social justice. Issues of social justice are often impacted by teacher evaluations. While student achievement is a key value or goal of schools and good teaching should be clearly identified, instruments to identify good teaching are often lacking validity and do not take into account significant factors. The effects of the teacher evaluation systems must be evaluated in relation to its intended impacts on teaching and learning and social justice.

Student growth is often defined as **t**he average gain in student test scores from one year to the next. It compares the test performance of a group of students in one year with the test performance of the same group of students the year before. If all students are promoted normally, student growth measures compare the test performance of a group of students in one grade with the test performance of the same

group of students in the previous grade. The reality is that not all students are promoted normally, so equivalency by group by grade is fallacious and does not take into account social or cultural values or outcomes.

Systematic errors are present in attainment measures because schools serving low-achieving students are often destined to fail. This is because factors outside of the school's control that affect student learning are not taken into account in most of these measures. In growth measures, random errors are also present because when a student takes the first exam as the baseline for future progress. No one can be certain that he/she tested at his/her full knowledge level. Therefore, if all future growth is based on an inaccurate first test, then how can this measure be an accurate picture of real growth? Value-added assessment, a statistical process for looking at test score data that utilized additional factors, is one technique that researchers have been developing to identify effective and ineffective teachers and schools.

2. Teacher Evaluations and Considerations for Assessment

One of the most important questions about value-added assessment is whether the estimate obtained from a value-added model can actually be called a teacher effect. Some key questions need to be addressed about teaching practices and evaluations. What changes in teaching practices are reported by teachers and documented by observational measures and student ratings? What changes occur on high-stakes achievement tests compared to the baseline year, and are these effects confirmed by independent audit tests? What is the overall impact of these on social justice?

The general theory of action for test-based teacher evaluation systems holds that using student growth to measure teacher effectiveness will improve the quality of education provided to students and hence will improve student achievement. Value-added measures (VAM) is a measure of how much a student has learned from one point in time to the next, and is often used to describe a teacher's effectiveness on a group of students' academic growth from year to year [8]. Value-added uses a student's own academic performance as a basis for determining his or her academic growth and is not related to a student's socioeconomic status or other personal characteristics that typically confound achievement-based measures.

Value-added models purport to track the amount of value that a teacher adds to student learning from year to year [9]. Value-added information allows educators to assess their impact on student learning, and it can help initiate conversations about the efficacy of curriculum and instructional practices and programs. Value-added information also allows educators to better identify what is working well and areas for improvement to help individual students and groups of students. Above and beyond the estimates for summative evaluation, there is a wealth of diagnostic information being provided that can be appropriate for educators.

Proponents of test-based teacher evaluation argue that growth in student achievement is the ultimate criterion for judging teacher effectiveness. They believe that value-added modeling of test-score data can do a better job of identifying the best and worst teachers compared to current indicators and that these methods are sufficiently robust in accounting for initial student differences to provide actionable data [10]. People who oppose the use of VAM claim that neither standardized tests nor VAM's statistical methodology have sufficient validity for the high-stakes purpose of individual teacher evaluation and teacher pay [11].

Using value-added models in teacher evaluations has become the hot issue, with several states passing laws making this sort of measure of student achievement a significant factor in teacher evaluations--at least 50 percent in some states [12]. When it comes to improving and assessing teacher effectiveness, conditions in the school are often not taken into account. Some of these conditions include class sizes, school cultures, student mobility, area economic issues, language issues, supportive school cultures, access to needed materials, and professional development opportunities for teachers, educational assistants, and administrators.

There is more to being an effective teacher than raising standardized test scores, yet test scores have gained widespread acceptance among the public as the key indicator of teacher performance. Value-added models attempt to isolate the impact a teacher has on students' achievement from other factors of interest, such as student characteristics [14]. It is important to consider the human side of teaching and learning as well as standardized test scores in measuring teacher effectiveness because teaching consists of classroom interactions among teachers and students, and teachers facilitate students' achievement of learning goals [13]. Communication plans for teacher evaluations and the process that is

being used, or planned, should take into account the process, the duration of the process, and plans regarding the utilization of the outcomes.

3. Examples of Evaluation Systems

The State of Minnesota in the United States passed legislation requiring all school districts to develop evaluations systems for principals by 2013-2014 and for all teachers by 2013-2014. Since spring 2013, the Minnesota Department of Education (MDE) has been partnering with 18 Minnesota school districts and charter schools to pilot Minnesota's example teacher evaluation model. The pilot is being implemented in order to inform improvements to the state model and to offer recommendations to all Minnesota districts as they prepare to fully implement teacher evaluation activities in school year 2014-15. The 18 pilot districts began implementing the model in fall 2013.

Sites that align their student learning goals with an overarching district or school goal will find that the goal-setting process is more clear and straight-forward, offering rich and timely feedback, and summative evaluations. They will also find that specialist teachers and non-teaching staff (e.g., nurses, counselors) who struggled much more than did classroom teachers to adapt the student learning goals to their situations, will be better served. The other items that are improvements are that teachers will view the evaluation process and this model as an effective form of professional development. Teachers and those who were involved as summative evaluators expressed hopes for increased collaboration through peer reviewer relationships, and sustainability of the model over time, was a repeated concern of teachers and summative evaluators – especially with the factors of coast and time. Recommendations from teachers and summative evaluators **a**re to provide samples of completed individual growth and development plan forms, student learning goals forms, and points of contact documentation; and to clarify the relationship between goals on the individual growth and development plan and student learning goals.

In the Minneapolis Public Schools, as one example, teachers are evaluated based on the use of three different tools: classroom observations, a student survey and student achievement data. Think about that from the perspective of a kindergarten or early elementary school teacher. There are four classroom observations per year. The teacher receives points based on standardized criteria and as reported by teachers, the feedback is generally helpful. But these observations also involve the observer walking up to students and asking what they are doing. All students, even my 5-7 year-olds, who may have just started school, or are in the early grades, get asked this question. The student is supposed to regurgitate the "I can" statement that correlates to "Focused Instruction." This is very difficult, if not impossible for some very young students.

The second tool is student surveys, which are typically administered by the classroom teachers once each year. The students have to circle their responses (even if they can't read) to questions about their teacher and school. When you have students that are homeless, highly mobile, in high poverty situations, or come from homes where violence is constantly present, under no circumstances does this student survey make sense for them or to them.

The third item used to evaluate teachers are student achievement data. Two to three times a year, students are pulled out of classrooms and tested by a stranger from the district. When the tester asks the students to go into a separate room and gives them a test, many of them shut down. It can be very intimidating to them since they do not know the tester and they may not trust strangers. Some of the students are asked to take this test in the middle of breakfast; others are tested right after recess. The inconsistency of when the students are tested creates a test that isn't being measured consistently or accurately.

If there are issues with the retention rate of teachers it will not go up unless there is more incentive to stay — and more assistance to attempt to give the students an even chance. As an example, the Minneapolis Public Schools replaced 200 teachers this past year and does not have the resources or personnel to effectively mentor them. The Minneapolis Public Schools employs 3,789 teachers and has 1,647 instructional support personnel. They have 35, 356 students of which there are: 11,439 white students, 13,109 African American students, 6,766 Hispanic students, 2,526 Asian students, 1,479 American Indian students, and 37 Pacific Islander students. Of that group 18% are in special education, are eligible for 65% are free/reduced price meals, and 24% are English language learners. As you can tell, some students are in more than one category.

4. Potential Issues

A potential problem that needs to be addressed is the result of the unequal allocation of principals and teachers within schools as well as additional demands placed on teachers and principals in more disadvantaged schools. This may impede their abilities to implement necessary reforms to improve academic achievement of students and account for a lack of fairness in teacher evaluations. For example, if higher-quality principals and teachers are concentrated in higher-achieving, lower-poverty schools, it should not be surprising that a program that relies on high-quality principals and teachers has larger effects in these schools. In addition, less-advantaged schools with, on average, harder-to-serve student populations, may require additional supports for these kinds of interventions to generate improvements in student learning similar to those of more-advantaged schools.

A central question that must be addressed is how do we fairly account for the effect of the many types of diversity in classrooms? There is an extensive amount of attributes beyond cultural differences that must be taken into account. In a single environment, learners and teachers themselves vary in beliefs, attitudes, perceptions, self-efficacy, motivation, learning styles, cultural influences, and demographics or social identities (e.g., sex, sexual orientation, ethnicity, ability/disability, socio-economic status, religion/spirituality, etc.). If you look at the many levels of diversity just noted, several (e.g., attitudes, motivation, self-efficacy, etc.) are typically not considered in calculations since data are not collected for them.

Value-added models cannot fully control for variables because neither teachers nor their students are randomly assigned to either schools or classes, making it difficult to separate a teacher's impact on students from other non-observable measures, such as a student's motivation or help at home [14]. The most significant finding from a Rand Corporation investigation into value-added models is that because such models might not control for all variables of interest, student achievement can never be shown conclusively to be due to individual teacher effectiveness [15].

Another significant issue is the use of Common Core Standards and tests. Common Core Standards do not tap into, nor do they assist in, reducing the achievement gap [16]. While there is a need for consistency in measures and Common Core Standards seek to provide that consistency, the real crisis is not a lack of common core standards but racism and poverty. Racism and poverty are typically not being taken into account in the teacher assessment processes currently being utilized. By privileging one way of being literate and one way of making sense of texts, she says, the standards fail to recognize and value those students who embody various "funds of knowledge" reflecting diverse families and neighborhoods [16].

Teachers make a difference and there is a link between teacher effectiveness and student learning [17]. VAM can be useful and the whole point of VAM is to create a more level playing field in order to make more fair comparisons among teachers. Policymakers and educators understand that raw achievement test scores tend to rank schools by the socio-economic status of the students served and are not fair, or consistent measures of teacher success. The very name, value-added, reflects the desire to isolate the unique contribution of schools or teachers to achievement outcomes.

Evaluation systems should take into account school and student variables and give more meaning to the career and compensation ladder for teachers by helping them to engage proactively in valuable professional development opportunities. In order to provide important feedback, value-added measures should be used only when there is a sufficient sample size and multiple years of data that take into account significant factors noted in this research report. However, many teachers have few students linked to them for whom data is available for both prior-year and current-year achievement. Other students who are mobile may have spent only a short time in a given teacher's classroom. Both of these are sources of considerable error. Year-to-year instability in teacher rankings is also very high. Many experts suggest that there should be at least 50 students (who have been with the teacher for a large majority of the year in each case) and at least 3 years of data to use in estimating a value-added score. Even with these considerations, it is important to recognize that multiple years of data may mask the year-to-year instability of scores, but do not eliminate the causes of such instability, which may often include the composition of classes that teachers teach.

5. Conclusion

The whole point of taking into account the value added measures is to create a more level playing field in order to make more fair comparisons among teachers. Policymakers and educators understand that

that raw achievement test scores tend to rank schools by the socio-economic status of the students served and are not fair, or consistent measures of teacher success. The very name, value-added, reflects the desire to isolate the unique contribution of schools or teachers to achievement outcomes. It would be more accurate to measure student growth over a specified period of time and make allowances for highly mobile students and those who may not have started school at the beginning of the school year.

The use of any value-added measure should take into account characteristics of the students and the context that affect student achievement gains. Such factors include parent education, special needs of students (English learner and special education status, poverty, homelessness), student attendance, and classroom composition, in addition to the individual student's prior achievement [18]. In particular, studies show that classroom composition greatly affects teachers' value-added scores. This information should be taken into account both in the models and in the overall analysis of information for the ultimate evaluation judgment. Other factors that may make a significant difference include class size, the quality and availability of curriculum materials, whether students also receive tutoring or related instruction from another teacher, etc. If these factors are not accounted for in the value-added model, they should be accounted for in the overall evaluation of a teacher.

Value-added measures should be used only when there is a sufficient sample size and multiple years of data. Studies find that many teachers have few students linked to them for whom data is available for both prior-year and current-year achievement. Other students who are mobile may have spent only a short time in a given teacher's classroom. Both of these are sources of considerable error. Year-to-year instability in teacher rankings is also very high. Many experts suggest that there should be at least 50 students (who have been with the teacher for a large majority of the year in each case) and at least 3 years of data to use in estimating a value-added score. Even with these considerations, it is important to recognize that multiple years of data may mask the year-to-year instability of scores, but do not eliminate the causes of such instability, which may often include the composition of classes that teachers teach.

The validity of teacher effectiveness ratings in any given state or district or school will depend on several factors such as the particular achievement measures used to assess the outcomes of learning, the adequacy of prior achievement data, the assignment of students to classrooms, the concurrent effects of other learning resources, the particular Value Added Measures (VAM) specifications, the quality of observational and other measures of effectiveness used in the system, and on the judgments involved in weighing evidence from multiple measures. At best, existing research offers insights about the potential threats to validity that need to be addressed in order to create systems for analysis and evaluation that are more fair and take into account social and cultural variables for social justice.

The end-of-year test scores do not show how much students learned that year in that class with that teacher. Measures that take into account where students started are an improvement however, such measures of growth are only a starting point. Making judgments about individual teachers requires sophisticated analyses to sort out how much growth may be caused by the teacher and how much is caused by other factors. For example, students who are frequently absent tend to have lower scores regardless of the quality of their teacher, so it is vital to take into account how many school days students are present. Thus, to be fair and to provide trustworthy estimates of teacher effectiveness, value-added measures require complicated formulas that take into account as many influences on student achievement as possible as noted in this article.

While much of the debate over evaluation systems centers on their use of student test-score data to measure a teacher's "value added" to student learning, classroom observations remain critically important. Most teachers work in grades or subjects in which standardized tests are not administered and therefore will not have a value-added score. Even when students' test scores are available, classroom observations may capture dimensions of teachers' performance that are important but not reflected in those scores. Finally, value-added scores on their own do not tell teachers how they might improve their practice and thereby raise student achievement.

If the principal, who is typically the one to evaluate teachers, changes their role from pure evaluation to a dual role in which, by incorporating instructional coaching, they served as both evaluator and formative assessor of a teacher's instructional practice. It seems reasonable to expect that more-able principals could make this transition more effectively than less-able principals. The principals could then provide formative and summative feedback and recommend appropriate professional development for

the teacher. A very similar argument can be made for the demands that the new evaluation process placed on teachers. More-capable teachers are likely more able to incorporate principal feedback and assessment into their instructional practice. Improvements are needed in how classroom observations are measured if they are to carry the weight they are assigned in teacher evaluation. The report's authors make specific, evidence-based recommendations aimed at improving the fairness and accuracy of teacher evaluation systems. [19] Key findings and resulting recommendations include the following ideas.

Under current teacher evaluation systems, it is hard for a teacher who doesn't have top students to get a top rating. Teachers with students with higher incoming achievement levels receive classroom observation scores that are higher on average than those received by teachers whose incoming students are at lower achievement levels, and districts do not have processes in place to address this bias. Adjusting teacher observation scores based on student demographics is a straightforward fix to this problem. Such an adjustment for the makeup of the class is already factored into teachers' value-added scores and it should be factored into classroom observation scores as well.

The reliability of both value-added measures and demographic-adjusted teacher evaluation scores is partially dependent on sample size. These measures will be less reliable and valid when calculated in small districts than in large districts. Thus, states should provide prediction weights based on statewide data for individual districts to use when calculating teacher evaluation scores.

Observations conducted by outside observers are more valid than observations conducted by school administrators since there may be a built in bias, whether conscious or unconscious, by an administrator in the system. At least one observation of a teacher each year should be conducted by a trained observer from outside the teacher's school, preferably one who does not have substantial prior knowledge of the teacher being observed.

The inclusion of a school value-added component in teachers' evaluation scores negatively impacts good teachers in bad schools and positively impacts bad teachers in good schools. This measure should be eliminated or reduced or revised for use in teacher evaluation systems. Effective evaluation systems should be based on professional teaching standards, include multi-faceted evidence, include the use of student demographic factors, use knowledgeable evaluators and a team, use evaluations that contain useful feedback connected to professional development, value and encourage teacher collaboration, use expert teachers as part of the assistance and review process for new teachers and those needing extra assistance, include a panel of teachers and administrators who oversee the evaluation process, and are continually evaluated and re-designed to meet current needs and demographic changes in the student population.

While I think we all agree that evaluation systems have an important role to play in making every teacher more effective and in examining student growth, this is a difficult process. Students need to achieve at an acceptable rate, taking into effect the variables noted in this paper is a difficult process, and staff development needs to address all of the elements of the evaluation system. If evaluation systems go beyond carrot-and stick diagnostics of "good" and "bad" teachers, and instead are used as systems to support professional development, teachers and unions will be much more willing to support evaluation reforms. Well-designed assessments that are administered in both a formative and summative manner and that are aligned with curricula and take into account student and cultural variables and are focused on higher-order skills will be very useful. If the systems have timely reporting of results they can be useful tools to support effective teaching and student achievement in every subject and grade.

References

[1] Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

[2] Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247–256.

[3] Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement* (No. R11-0435-02-001-97). Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

[4] Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

[5] Rice, J. K. (2003). *Teacher quality: Understanding the effectiveness of teacher attributes*. Washington, DC: Economic Policy Institute.

[6] Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73*(1), 89–122.

[7] 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.

[8] Marzano, R., & Toth, M. (2013). *Teacher evaluation that makes a difference*. Alexandria, VA: Association for Supervision and Curriculum Development.

[9] Amerin-Beardsley, A., & Collins, C. (2012). The SAS Education Value-Added Assessment System (SASS EVASS) in the Houston Independent School District (HISI): Intended and unintended consequences. *Education Policy Analysis Archives*, *20*(12), 1-27.

[10] Gordon, R., Kane, T. J., & Staiger, D. O. (2006, April). *Identifying effective teachers using performance on the job*. Washington, DC: The Brookings Institution.

[11] Braun, H., Chudowsky, N., & Koenig, J. (Eds.). (2010). *Getting value out of value-added: Report of a workshop*. Washington, DC: The National Academies Press.

[12] Institute for Competitive Workforce. (2011, January). *In focus: A look into teacher effectiveness*. Washington, DC: US Chamber of Commerce.

[13] Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning* (Vol. 1, pp. 371-404).

[14] Hull, J. (2011). *Building a better education system: At a glance*. Alexandria, VA: Center for Public Education.

[15] McCaffrey, D., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: Rand Corp.

[16] Compton-Lilly, C. (2014). *Common core standards do not account for diversity*. Retrieved May 1, 2014 from http://www.wcer.wisc.edu/news/coverStories/2014/common_core_standards.php

[17] Tucker, P., & Stronge, J. (2005). *Linking and student learning*. Alexandria, VA: Association for Supervision and Curriculum Development.

[18] Darling-Hammond, L. (2012*). Creating a comprehensive system for evaluating and supporting effective teaching*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

[19] Whitehurst, G., Chingos, M., & Lindquist, K. (2014). Evaluating teachers with classroom observations: Lesson *learned in four districts*. Washington, DC: Brookings Institute.