

# **Teacher Evaluations: What Are We Really Evaluating?**

**By Brad Suder**

## **Introduction**

Since the initiation of No Child Left Behind in 2002, the push for teacher accountability has continuously increased. We now face a pivotal period in education that is attempting to reform the system all together. One area where this change is greatly taking hold is in the realm of teacher evaluations. Current trends in evaluation are creating many questions from many stakeholders in the equation. What do we want students to learn? What is the goal of education? How should teachers be held accountable for student learning? How do we measure what students are learning and how effective teachers are? These are just a few of the many questions that arise from such a hot topic.

With the rapidly growing support for increased teacher accountability, the push for more standardized testing has resulted. Many people making decisions in education (typically not educators) feel these tests are the best way to evaluate teacher effectiveness. I believe that this form of evaluation is first and foremost a flawed model to evaluate teacher effectiveness and student learning. I also believe this movement education will have detrimental effects on students, teachers, and the education system in its entirety.

I will first explain popular opinions on the need for reform of teacher evaluation. Second, I will explain some of the new models being implemented and proposed. Then I will explain the oncoming model that will be implemented in Ohio. Lastly I will express what I believe would be the best means of assessing teachers.

## **Reasons for Change**

There are many reasons people would like to see current evaluation systems change. Simply put, many people feel they are an ineffective formality. As Almy (2011) states, “We judge teachers based on a small number of fleeting classroom observations. These evaluations do little to help teachers understand the impact they have upon the learning of their students. Moreover, because so many evaluations lack detail and clarity, they provide teachers little information about what to improve or how to get better” (p. 2). This is an outlook that is echoed by many stakeholders both in and out of education. “The typical teacher evaluation in public education consists of a single, fleeting classroom visit by a principal or other building administrator untrained in evaluation who wields a checklist of classroom conditions and teacher behaviors that often don't focus directly on the quality of instruction” (Toch, 2008, p. 32). Weisberg stated, “in a study of evaluation programs in twelve school districts, The New Teacher Project found that less than 1 percent of all teachers are rated as unsatisfactory” (as cited by Papay, 2012, p. 126). I know from my own experiences with evaluations, this is not far from the truth. I've had several post observation follow-ups where I was told “everything was great.” While I would like to think everything I do is great, I know that as a fourth year teacher, I am far from perfect and have ample room to improve. Many people also push for reform to differentiate between high and low quality teachers. In some cases this is for the means of rewarding the good and removing the bad, and in other cases it is just to acknowledge what is working and to seek improvement upon what is not. In regard to the former, I think we must limit how much stock we put into this and give our attention to the latter. This will truly bring a school together to strengthen it.

## **New Models**

The two current trends in teacher evaluation that are establishing a large presence in schools across the country are standards based classroom observation models and value added models (Papay, 2012, p. 124). These two models take very different approaches to evaluating the effectiveness of a teacher. Standards based models rely heavily upon classroom observations and a specific set of standards with which to base these observations. One widely used set of standards is the Danielson model. Charlotte Danielson created a set of standards based on four main categories with 22 themes and 77 skills (Toch, 2008, p.33). This is clearly a very detailed rubric, which is good, but it is also so large that it makes it nearly impossible to evaluate a teacher on all of these measures in a small number of evaluations. Many of these new evaluation systems push to observe teachers anywhere from two to six times in a school year. While that is better than some districts that fail to observe a teacher more than one time, I still don't feel that four evaluations in a year can paint a complete picture of a teacher.

One specific standards based model is called TAP (Teacher Advancement Program). It requires three classroom observations by different evaluators based upon a modified version of the Danielson model. These evaluators are mostly master teachers who are trained to conduct classroom observations. Teachers are then observed by master teachers who specialize in their content area (Toch, 2008). I particularly like this idea. Having an evaluator who understands your content and best ways to teach it could provide better feedback and a more accurate evaluation. A colleague of mine who teaches math has said that many times the people performing her observations do not understand how to do any of the math she is teaching and just assume she's doing a nice job.

There are many other ideas that have been implemented in standards based models. One model involves evaluating teachers' portfolios that would encompass lesson plans, student work, etc... While this is a nice idea and is easier on the evaluator as far as finding time to observe in classrooms, it fails to observe many of the vital aspects of teaching, such as the classroom environment, classroom management, student engagement, teacher-student rapport, etc... Another model includes forming teams of teachers to evaluate their colleagues as many as six times in a semester. I think this frequency is great and could really provide some useful feedback. One of the obvious restraints of this is time. It would be very costly and time consuming to have teachers in other teachers rooms this often. This model actually uses these teacher observations as a means of retaining and removing teachers, which I have great concern about. That would really create a separation among teachers and your colleagues evaluating you could be turned into your enemy. A final model of standards based evaluation included creating a committee of master teachers who are removed from the classroom for several years to take on the job of evaluating. This is great to have teachers in this role and doing observations, but again it costs a lot of money to do so. It is essentially like creating a team of pseudo-administrators because the actual administrators don't have the time to do the job well. I applaud these schools for acknowledging that, but in this day and age where funding in education is more scarce than ever, I don't see this model as being a feasible option.

While I believe standards based evaluation models are a good way of evaluating teachers, the obvious drawbacks are that they are much more labor intensive and expensive. According to Toch (2008, p. 36), "At \$ 1,000 per teacher, it would cost \$3 billion each year to evaluate the 3 million teachers in the United States using a National Board-like portfolio or TAP's multiple evaluations/multiple evaluators model." Many argue that the cost is well worth it, because it would increase teacher effectiveness and therefore increase student learning (Papay, 2012; Toch,

2008). I happen to agree, especially in comparison to other methods that are still cost indulgent and seem to have more drawbacks. While standards based models could still have some of the same problems that existed in older observation models (teacher bias, inefficient evaluations, lack of quality feedback, etc...), if conducted well and held to a certain standard, they could really help all teachers to improve.

The other heavily trending system of teacher evaluation is the value-added model. “Value-added models enable researchers to use statistical methods to measure changes in student scores over time while considering student characteristics and other factors often found to influence achievement” (Darling-Hammond et al., 2012, p.8). This model looks at a student’s score in a subject on a standardized test and measures how much they grow each year and compare it so similar peers (“those with comparable performance histories and background characteristics”) (Almy, 2011, p. 4). The statistical methods used for this model are very complex and hard to understand for the average individual. While this model sounds great in theory since it is an ongoing measurement and it takes into account students’ backgrounds, there are numerous flaws. First, measures of teacher effectiveness are wildly inconsistent from year to year. “A study examining data from five school districts found, for example, that of teachers who scored in the bottom 20% of rankings in one year, only 20% to 30% had similar ratings the next year, while 25% to 45% of these teachers moved to the top part of the distribution, scoring well above average” (Darling-Hammond et al., 2012, p. 9).

A second problem lies in the students assigned to them. Many studies have found that teachers with higher numbers of students with special needs, English language learners, and students with a low socio-economic status typically have much lower value added scores, despite the model taking students’ backgrounds into consideration (Darling-Hammond et al., 2012, p.10). A third fault of the value-added model is that it is based upon students’ performance on a

standardized test. Basing the effectiveness of a teacher on one test given in a school year is simply ridiculous because of the many factors that could play a part in this.

First, not all schools are the same, therefore some students have more of an advantage due to the assignment of teachers to schools and the quality of schools. Secondly, what about the other factors that play a role in a student's success, for example, what if their parents got divorced shortly before the test, or they were very ill and missed the past month of school. Third, what kind of learning are these tests evaluating? "VAM-style measures may be influenced by how much the teacher emphasizes short-run test preparation" (Darling-Hammond et al. 2012, p.11). Evaluations models like value-added encourage teaching to the test and require students not to learn, but to memorize. "If student test-score growth itself does not reflect actual learning, then these estimates will not be valid for drawing inferences about teacher performance" (Papay, 2012, p.133). As Pinar (2012) exclaims:

Students are pressured to score higher on standardized test that allegedly measure education achievement. Never mind that test taking has never been – until now - conflated with education achievement (always before it was only one indicator among others and not necessarily the most significant one), and never mind that the "skills" these tests purport to measure are themselves abstractions severed from situations, reflecting the self-referential and enclosed world of standardization, academic versions of crossword puzzles. Driven by such self-enclosed rituals, education institutions devolve into cram schools, no longer about the world but, instead, about themselves, about those tests, apparently technical but altogether ideological, as students learn to process information without raising questions about that information or the process. (p. 53)

A fourth flaw in the value-added model is that different tests and different statistical models are used which greatly change the outcome of a teacher's value added score (Murnane and Steel, 2007, p26). If such a model is to be implemented, it is imperative that this becomes more consistent.

Simply put, value-added measures do not measure the kind of teaching and learning that I became a teacher for. They degrade the teaching profession, equating teachers to nothing more

than someone who spits out information, and students to nothing more than someone that can regurgitate it. Education is not designed for this, and contradictory to the aims of this business influenced model, it would be counter-productive to providing students with the long term skills that allow them to problem solve, think critically and creatively, and learn how to work and interact with others.

### **What's In Store for Ohio?**

The Ohio Teacher Evaluation System (OTES) which is supposed to be implemented in the 2014-15 school year is a dual system in which 50% is comprised of standards based teacher performance which will include a mid-year and final evaluation that will “consist of two formal observations of the teacher at least thirty minutes each in duration, as well as periodic classroom walkthroughs” (ODE, 2011). So essentially, four thirty minute observations will determine half of my effectiveness rating, which could be “ineffective, developing, proficient, or accomplished”. The other 50% of my evaluation will be based on value-added measures. It will be quite interesting to see how they manage to do this for every content area in schools.

Currently Reading and Math are the only subjects completely assessed. It will cost a great deal of money to create and grade tests for every other subject area. My rating merged of both parts, will determine what kind of professional growth plan I or someone else will develop. As a whole, I don't have a problem with the standards based portion, it is essentially what my school has been doing, but perhaps now will be with a little more accountability on the evaluator's part. I have a great number of grievances against the value-added portion, as many of the aforementioned weaknesses point out. This evaluation system as a whole will not only be costly, both in terms of spending time and money on classroom observations, but also in terms of creating and scoring standardized tests for all grades and content areas. All of this money could have been spent much more effectively, which would have led to copious amounts of learning, rather than

“cramming.” In addition to being wasteful with educational funding, it will also make teaching a competition, pitting teachers against one another and against the community when their scores are published in the newspaper. There is so much more behind establishing a teacher’s effectiveness and it sickens me to minimize our profession to a test.

### **So What Should We Do?**

As I’ve managed to tear down many of these proposed models, I’ve failed to mention what I think would be best. I believe that a standards based model is our best alternative, but there are many kinks that must be work out. First and foremost, it should be used primarily, “to improve instruction by developing teachers’ instructional capacity and effectiveness” (Papay, 2012, p.134). I also think it is impossible to expect the limited number of administrators in a school to perform all of these observations effectively. I think a system of teachers observing teachers should be executed to best provide useful feedback for classroom teachers. These classroom evaluations can be incorporated into teacher evaluations in addition to administrative observations. I think using a model like Danielson’s is a good idea and encompasses the entire role of the teacher (planning, instructional practices, classroom environment, and professional responsibilities) and not simply what they do during the half hour they are observed. I think creating improvement plans for teachers is a good idea, regardless of their observed effectiveness, because all teachers should be continuously working to improve. It should not be given or received as a punishment. We must be careful not to do this as a formality, either. We need to have follow-through and management of these plans. The goal of evaluating should not be to fire people. That said, I think tenure in schools is counterproductive to teacher improvement. I understand its original purpose, and it is obviously a secure feeling to have, but it takes away the motivation to get better in many circumstances. It shouldn’t be removed to

simply fire people, but to encourage people to improve their practices and not to have a safety net if they choose not to.

In conclusion, teacher evaluation is a very complex subject with no easy solution. There are obvious drawbacks to any possible model. I think the goal now is to establish a system that has the fewest amount of faults and will improve teaching and student learning the most. I don't feel that current systems or oncoming value-added models are successful in truly helping teachers to become better teachers, and they certainly don't encourage high levels of learning. All stakeholders involved in the process must take time to evaluate the purpose of education and what they truly wish to achieve through their ideal evaluation system. If we aren't seeking to help students, then we aren't doing what is right.

### References

- Almy, S., & Education, T. (2011). Fair to Everyone: Building the Balanced Teacher Evaluations that Educators and Students Deserve. Teacher Quality. *Education Trust*.
- Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating Teacher Evaluation. *Phi Delta Kappan*, 93(6), 8-15.
- Evaluation of Professional Staff. (2011). Paper presented by multiple authors. Dayton, Ohio: ODE Race to the Top Regional Conference Meeting.
- Murnane, R. J., & Steele, J. L. (2007). What Is the Problem? The Challenge of Providing Effective Teachers for All Children. *Future Of Children*, 17(1), 15-43.
- Papay, J. P. (2012). Refocusing the Debate: Assessing the Purposes and Tools of Teacher Evaluation. *Harvard Educational Review*, 82(1), 123-141.
- Pinar, W. (2011). *What is curriculum theory?*. (2nd ed.). New York, NY: Routledge.
- Toch, T. (2008). Fixing Teacher Evaluation. *Educational Leadership*, 66(2), 32-37.