# National Board Certified Teachers' Perspectives on Using Growth Measures of Student Learning for Teacher Evaluation

James H. McMillan
Virginia Commonwealth University

March 10, 2015

## Introduction

Policies for the improvement of student learning in America have most recently focused on the impact of teachers on student achievement, and high-stakes teacher evaluation has become the primary mechanism for enhancing teacher effectiveness (Darling-Hammond, 2013; Lavigne, 2014). Many traditional procedures for evaluating teachers have been deemed ineffective, most particularly scant administrator observation of teaching, cursory reviews of goals, lessons plans and activities, brief and unhelpful feedback, standardization, and lack of consistency among evaluators (Danielson, 2011). In addition, there is ample evidence about what should be avoided. For example, we know that reforms in teacher evaluation will be ineffective and even damaging if they are unmanageable, unreliable, reinforce teacher isolationism and competiveness, inhibit collegial activity, and rank teachers on the basis of accountability test scores or other measures (Darling-Hammond, 2013; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012). Despite this evidence there is continued interest in using accountability test scores to calculate value-added or other growth indices for individual teachers. This is based in part on research that has shown that specific teacher classroom practices correlate with end-of-course proficiency tests (Kane, Taylor, Tyler, & Wooten, 2011), and

that results predict to some extent student learning and other outcomes (Chetty, Freidman, & Rockoff, 2011; Harris, 2011).

It is clear that effective teacher evaluation needs to include measures of student learning, along with more thorough administrator observation, review of instructional goals and materials, peer review, and student feedback (Marshall, 2012; Marzano, Frontier, & Livingston, 2011; Marzano & Toth, 2013; Stronge, 2010). However, the use of student test scores as a way to document student learning is fraught with difficulties, especially when student accountability test results are used for judging value-added learning (Marshall, 2012), an approach that has significant negative consequences (Konstantopoulos, 2014; Lavigne, 2014; Lavigne & Good, 2014). While it is not possible here to explicate all the technical and practical issues that show how value-added approaches are deleterious and result in serious errors, it is important to note that its the usefulness for improving teacher evaluation is limited at best and dangerously inaccurate at worst (Baker et al., 2010; Berliner, 2014; Papy, 2012).

Given the controversies surrounding the use of value-added and other growth measures, school divisions are turning to pretest-posttest designs for segments of learning, typically measuring students' gains in achievement over six or eight weeks, and student goal-setting. Like value-added models, however, there are significant weaknesses with this approachs. Often such measures lack validity and reliability, and still do not provide results that help teachers improve their instruction or student learning (Abrams, Varier, & McMillan, 2012).

There is universal agreement that evidence of student learning should be incorporated as an important and substantial aspect of teacher evaluation. What is now at

issue is the nature of the measures of student achievement that most fairly and accurately indicate what teachers have been responsible for in students' learning progress, and how use of these measures ensures teacher quality and also improves instruction and subsequent student learning. What is most glaringly absent in determining this, due to the allure of using quantitative indices, are voices of teachers about what types of measures of learning would be most valid, fair, and helpful in meeting these needs. In particular, there has been very little consideration of how classroom assessments (e.g., tests, quizzes, papers, projects, and other day-to-day assessments that teachers construct, score, and grade) could be accurate indicators of how teachers influence student learning.

## Purpose

The purpose of this investigation was to gather opinions from National Board Certified Teachers (NBCT) about using student test scores for teacher evaluation, and the possible use of results from classroom assessments as evidence of student learning. The aim was to listen to teachers who had been recognized for their outstanding expertise, who had demonstrated a long-term commitment to the improvement of their teaching, and who had been positively evaluated by others based on the submission of multiple types of evidence, including evidence of student learning. The hope was this group of teachers, because of their established teaching competence and professional involvement, would provide particularly helpful perspectives about the benefits of and problems with using evidence of student learning in evaluating teachers.

## Methodology

Four focus groups of NBCTs were held in May, 2013. The teachers were contacted by email from a list of area NBCTs from a large metropolitan region in a southeast state,

and volunteered to attend one of the four meetings. A total of 20 teachers from four school districts attended the sessions (8 high school, 3 middle school, 7 elementary, and 2 specialists; 2 males). The teachers averaged 19 years of teaching experience. Two teachers were African American, one was Asian/Indian, and 17 were White. A variety of subjects was represented (7 math, 5 reading/language arts, 1 science, 1 art, 1 government, and 5 multiple subjects). The researcher met with each group for approximately one hour. Each group was assured that their responses would be anonymous. All agreed to having the discussions audio-taped. The interview questions focused both on general teacher evaluation procedures based on measures of student achievement as well as the use of classroom assessments for generating evidence of student learning. The researcher also took notes during the discussions. A semi-structured set of questions was used by the interviewer for each group, first focusing on opinions of current teacher evaluation efforts, then asking about the use of measures of student learning, both standardized and classroom assessments. The audiotapes were analyzed and coded, using constant comparison and identification of negative cases. Along with the researcher notes, the coded responses were used to generate themes, areas of agreement, and recommendations for instituting effective teacher evaluation procedures. A summary of the findings was sent to the participants as a member check, with no corrections or needed revisions indicated.

## Findings

The most prominent types of measures of student achievement used to document growth in learning for evaluation for these teachers included end-of-course accountability test scores, in courses required to be tested, the use of interim or benchmark test results for their class, student goal-setting, or pretest-posttest results covering several weeks of

course content. The overall attitude of the teachers toward the use of these measures for teacher evaluation was mixed, though more negative than positive. While a majority indicated that they thought the current approaches (e.g., using pretest-posttests and student goal-setting) were in the main ineffective and bothersome, some thought that there were important potential benefits. The negative opinions were based primarily on their belief that the current approach was not valid because it could be "gamed," did not include all relevant contextual and student characteristic influences on learning, did not reflect important learning gains, and demanded even more "administrative" duties and record-keeping that took time away from preparation and actual instruction. It was clear that, in this sense, the negatives outweighed the potential positive benefits. They understood that test scores or other measures of student learning needed to be included in teacher evaluation, just that what was currently being done was not valid and very inefficient.

On the positive side, some teachers thought that because the system increased the emphasis on student learning there was a potential of providing helpful information to balance generally unreliable administrator observations and ratings of instructional materials. A few teachers also thought it was helpful to emphasize data-driven decision making and use of the results of student assessments to drive instruction (though more for other teachers they know about in the schools, not so much themselves). The teachers reacted to teacher evaluation in ways similar to testing for student accountability – the idea of including indicators of student learning makes sense logically, but methods and procedures must be fair, reflect what teachers and schools can be responsible for, and have positive consequences. Based on their comments, it is reasonable to conclude that teacher

evaluation systems that include measures of student learning need further development to meet these criteria. Here are two comments from teachers that illustrate these points:

- I can see both sides of the fence and it has built awareness for myself and going back to why I do assessment, and how to increase rigor.
- I'm of two minds. I think it's a great opportunity for us to really take charge of this profession. Take it back from the bureaucrats and say, this is what teacher evaluation should look like. It should make us better teachers, and this is what it should look like.

Beyond these general reactions the opinions of the teachers fell into six major categories, with some subcategories. The six categories included record-keeping and efficiency, contextual influences on learning, accuracy of results and gaming, goal-setting, consequences, and use of classroom assessments. Each of these will be discussed with direct quote illustrations that capture the nature and tone of teacher comments.

**Record-Keeping and Efficiency**

Perhaps the most salient and significant finding from these interviews was that teachers were resentful of what they see as largely ineffective record-keeping procedures that take valuable time away from instruction. The burdensome nature of yet more demands of teacher time was clearly articulated. While some participants saw some value to what was being required in documenting students' goal attainments, most were very negative about current time-intensive practices that were, in the main, not helpful in improving student learning or motivation. Many indicated that these demands have resulted in frustration, stress, and lower teacher morale. Valuable instructional time has been taken-away, and teachers spend more out-of-class time on record keeping and less on

providing feedback to students and preparing instructional materials. The tone of the following comments illustrates low teacher morale as well as significant concerns:

- The time that it [pretest-posttest procedures] took from me, the investment and the energy and everything, it has created a stress that is not needed because over the decade …. more and more and more is being expected and thrown upon us and there's no help, and there's no end in sight. And no matter what, you're damned if you do, and you're damned if you don't.
- I have found that the evaluation process [documenting pretest-posttest gains for all students] has become incredibly overwhelming.
- I don't love it [teaching] anymore. I shouldn't say that, but I don't. I used to absolutely love this job. I loved getting up. I love teaching the kids, but it's not that anymore. It's a paper work process.
- I don't know how [made-up pre/posttests are] anything other than just making work for people that doesn't help kids learn.
- Once again, then we're taking more time out of the classroom and more time to do this kind of crazy work.
- It's more and more the expectation to do more for the same pay and it's creating a very contentious friction within, at least from some of my colleagues what I've seen, is just this resentment of you want me to do more and it's only going to take five more minutes.

**Contextual Influences on Learning**

There was universal agreement that any system of teacher evaluation that incorporates measures of student learning must recognize that many factors beyond the control of the teacher impact student motivation and learning, and these factors must be

included in the interpretation of results from assessments of student learning. Student motivation to provide serious effort in completing the assessments was often cited as a problem, invalidating results. Some teachers cited students who simply did not care about demonstrating their knowledge or learning, and they were frustrated to the extent that results from these students would be used to evaluate their effectiveness. Others cited the unique characteristics of some classes, such as honors and AP classes (e.g., if students show very high scores on the pretest, how can you show improvement)? To provide a fair evaluation, these teachers clearly thought that student ability and motivation need to be considered in the interpretation of assessments of students learning. Often factors outside of school, such as home circumstances, parenting, neighborhood dynamics, or personal issues were very real and influential impacts. The teachers did not mind being evaluated on the basis of student learning as long as these factors comprise the context for interpreting the meaning of the results. Teachers voiced the following:

- We're dealing with outside variables having an influence upon how well a student is going to do. And I think that certainly is a factor that has to be taken into consideration.
- You can't just, when you're measuring like student growth, there's other, so many factors involved, you know, when we do research we're just changing one thing. When you're looking at a student and a classroom, there's so many other factors involved.
- I will come in my classroom and the heads are sometimes on the desk, they don't care.
- You know, I've got a kid that might show a lot of growth, but he may have lost a mother in the middle of the year, I mean, there's just so many, I'm looking at this from like a scientific view, there's so many other factors involved in a child.

- One of the invalidities that needs to be recognized is where it's the outside circumstance. I had a student who failed his SOL. He failed this thing four times in the test and I looked at him and said, why didn't you do well? And he said, simple, I knew I could take it a second time and because I didn't pass the first time, I get to get out of class so that way I can go ahead and be remediated.
- We cannot always use tests and grades to measure a teacher's academic success or academic progress when we're taking out the factor of human motivation.
- Now you give them a test, and they're probably not going to do too well on it because they don't care. They don't, I have one child who walks into class and all he brings with him is his cell phone. That's it. He doesn't have a backpack, he will not carry a backpack, he will not bring in a pencil or a piece of paper. He always has tomorrow. He has no motivation to do well.

An additional important contextual factor is the level of class that is taught. Teachers see different dynamics in student learning and expectations, and these differences need to be taken into account:

- I see so many teachers in this building who teach the lower level kids that don't get the recognition that they are excellent teachers whereas there are some higher level teachers who are wonderful too, but there are also some higher level teachers that have it real easy.
- My goal that I have for my GT class is way higher than the goal that I have for my on-grade level kids because … I want to challenge them, although I am challenging all my students, but my ones that I know can handle it, I'm pushing them even further.

**Accuracy of Results and Gaming**

A clear finding was that there is much "gaming" of the current pretest-posttest approach that use test score results, and goal-setting, to measure student learning to assure positive results. This included setting low goals and constructing tests that would show gains. Some of this is attributed to the high-stakes associated with the results. As one teacher indicated:

"My children's health and dental insurance rides on my back."

Nearly all of these teachers knew other teachers who manipulated assessments, scores or student goals to ensure evidence of student growth. For example, some teachers indicated using zeros on pretests that students simply failed to complete, using the zeros to calculate gain scores. This is at worst unethical teacher behavior, but one that is encouraged by the high-stakes now associated with teacher evaluation. Here is how two teachers explained how they "gamed the gains:"

- I'm not trying to sound unethical, but if you look at my data and you look at my students' names, the main reason why my students did so well on their posttest is because the majority of my students didn't do the pretest. They refused to write the essay. And what I think we're going to run into is, I mean, for me, my data is 100% accurate.
- I teach geometry, so when I give them the pretest, of course, they don't know anything. So the scores were anywhere between zero to ten or something … Of course, I knew when I was grading my posttests the grades would look higher because I would give them partial credit here and all that. But that was something for the administration… you can manipulate those answers giving them partial credit and everything.

**Goal-setting**

A significant portion of the evidence of student learning relied on teacher goal-setting. Many of the teachers interviewed knew others who would create learning goals that could easily be obtained. They would essential lower goals so that required high percentages of students would show improvement. This was motivated largely by the higher stakes associated with having a positive evaluation. In essence, the subjectivity and lack of standardization in setting goals resulted in faulty conclusions about student learning:

- I could set low standards for myself and end up being a great teacher.
- Some teachers are creating data so that it fits into the goal…they are creating a goal that will show that the students have succeeded.
- I don't think it's really making a difference because of what I have noticed at my school, teachers are writing goals that know that they can meet.
- They're going to write goals that they know they can meet.
- It is possible, especially with the geometry teacher, for me to give an assessment that every kid could pass on the first day, just the way I word it, and the vocabulary I use for the way I do.
- The pre-test and the post-test, I'm a Math teacher … we can manipulate those results.
- Some teachers are lowering it [goals] because they needed to get credentials and needed to pass.

Some of the teachers were clearly concerned about learning goals that were not part of the evaluation system. Several cited examples of significant student learning that were not captured by mandated tests, especially those used in a pretest-posttest design. At the

same time, teachers were concerned that many teachers in their schools "teach what is tested," narrowing what is emphasized to primarily what is on the tests. There was a clear consensus that it is not reasonable to evaluate students' academic progress on the basis of differences between pre and posttest scores from a single assessment, due primarily to the quality of the assessments and the fact that much significant learning was not captured. These teachers voiced the following:

- I'm still not convinced that measuring student progress is actually measuring the quality of teaching or the quality of teachers in our building. I'm thinking, what I'm seeing happening is the morale is down terribly.
- [The kids] even see it as a completely useless exercise. They knew they were learning. They're kids, but you can see their progress over a year, you can see them stretching or whatever… these couple of tests are completely a waste of time.
- For years I've put on a math night, it takes hours and hours and hours and hours to do it. And I love doing it and the children love it. There are 300 kids that come. Did it have an impact on my evaluation? No.
- And sometimes it's not the final assessment that is the most meaningful thing for the students.

**Consequences**

This sample of teachers clearly thought that the current methods of measuring student learning had negative consequences for themselves as well as their students. As previously mentioned, the record-keeping requirements were characterized as burdensome, and some teachers set lower goals for student learning, possibly limiting student achievement. Consequences for students were also largely negative. They saw

decreased motivation, focusing inordinately on test preparation, and less on meaningful learning. Teachers made these illustrative comments:

- Lower goals for students.
- Well, I hate to say negative, but I feel like even in the self-contained population, we're starting to teach to the test.
- Right, I told my administrator, I answer to a higher power at the end of the day and I've got to do what's right and not what's necessarily easy. And yes that hard work in there that I've done, you said it was fabulous but yet I am only mediocre. And I'm going to stay that way [according to the system]. But am I mediocre to my children? That's what matters to me.
- I think it's had a negative impact on teacher morale. I think that teachers are dissatisfied with what's happening. I don't feel like we are respected for what we're doing.
- I don't necessarily mind saying that you know, we have to be evaluated, the business world is evaluated. But our standards and means of evaluation are deadening the education system.
- Well, my daughter came home upset early in the year, she's like they're testing us on stuff we've never even gone over. And then they're being told to take it seriously because it's going to count.
- So the kids are freaking out because they're like we're getting tested on something we've never been told how to do in the first place.
- So it's having a negative impact on teacher morale.

The teachers also indicated that reliance on these assessments led to what they thought was an unfair categorization of the level of their performance, whether labeled "proficient"

or “advanced.” Here the conclusion about their teaching did not match with other evidence of student learning. This created anxiety, lower morale, and resignation to a system they viewed as unfair. Note the intensity of these comments:

- It bothers me tremendously because knowing that all of my students have made the goal that I set that we had gone over and they exceeded it, and I still didn't get exemplary and she couldn't answer my questions of what I could do to become exemplary. Couldn't answer it.
- I know that there are people who were expressing anxiety over being labeled proficient as opposed to exemplary.
- It is because you go on teaching 15, or 16 years and you have this level of proficiency and you are the same person but suddenly because of numbers you no longer can be, you're valued at the school level but beyond that nobody knows and they're not allowed to know it. And that’s not right.
- And these kids were taking 7 or 8 tests the first two weeks that basically were showing that they were not very good at these topics, not even getting to start out with good activities or whatever. It was completely against everything that that school has been about. And so I'm hugely upset by it.

**Use of Classroom Assessments**

The teachers were generally hesitant about the use of classroom assessments as high-stakes evidence of student learning to evaluate teaching, even though they viewed these types of assessments as reasonable and accurate indicators of what they have been responsible for with respect to student learning. It was clear that they had given little

thought to the use of classroom assessments for teacher evaluation, and were not loquacious about them, one way or the other.

The teachers thought that these types of assessments were more valid indicators of student learning than pretest-posttests, interim, or end-of-course tests, but were not very confident about how to structure the use of them so that it would not be a cumbersome process, easily manipulated to show positive results, nor have negative unintended consequences. Here are some illustrative comments:

- And the thing about classroom assessments is every teacher, they're writing their own classroom assessments. Some teachers write much more difficult classroom assessments than others. And you have to differentiate your classroom assessments for your class ability. I have a GT cluster and so mine are much more open ended, much more performance assessment based. So if we're using classroom assessments as our evaluation then for this 40%, it's going to be different for every teacher.
- Their performance in my class is a pretty good predictor of how well they'll do, I mean, it's not a perfect predictor, but it's pretty decent.
- It was more about like why we set up the things we did, and what we did to assist these students or guide them. It wasn't really about where they were, it was about how we used what we were measuring them with to advance them along. I think it was different.
- We're not looking at just a pretest assessment and a posttest assessment. We're looking at a collection of activities but once more, there's still that sense of you have to prove that this kid has learned.
- It certainly would be an incentive to people if they knew their assessments were being scrutinized, it could have an impact on what they were. I'm not saying that I would

change mine, but you know, I do think that people would think, would weigh what they were going to decide potentially differently.

For example, when suggestions about using student portfolios were made by the researcher the teachers had immediate concerns about the time and energy student portfolios would take, especially at the secondary level. Most agreed that only a sampling of students would be possible. How would those students be selected? How many students would be needed? What would be put in the portfolio? Would teachers focus more on the students selected and neglect other students? These types of questions were raised, without a consensus of how the portfolios could be effectively structured.

One interesting aspect of using classroom assessments focused on the possible use of peer reviewers to evaluate the evidence from the assessments. This was generally viewed positively, as long as the peer reviewers were trained and not in the same school. The teachers thought a peer review process would lessen the tendency to game classroom assessments.

## Discussion

The results of these interviews are best understood and utilized in the context of principles of teacher evaluation that are well established and supported. These principles include the following (Darling-Hammond, 2013; National Board for Professional Teaching Standards, 2011; Steele, Hamilton, & Stecher, 2010):

- Teacher evaluation systems must be *efficient and manageable*. A continuing significant issue in both accountability efforts and teacher evaluation is the introduction of burdensome procedures that take extensive time and resources away from instruction.

- Teacher evaluation systems should *not encourage isolationism, competitiveness among teachers, or ranking of teachers*. The nature of effective teaching is that it is a collaborative activity in which teachers provide mutual support and dialog about what practices best enhance student learning.
- The use of *unreliable or invalid measures of student achievement in teacher evaluation results in distrust, errors of judgment, and lack of support for conclusions*.
- Effective teacher evaluation *includes multiple sources of evidence of student learning*. There should not be a major reliance on single measures of student achievement.
- Teacher evaluation must take into account *differentiated instructional styles, the diversity of students, and the influence of contextual factors on learning*.
- Teachers should be evaluated on *what they can most directly be responsible for and influence*, what is in essence their contribution to student learning. They should be evaluated on what they actually teach, not on proxy measures that are less directly related to their instruction. Other influences on learning, such as class size, availability of tutoring, homelessness, poverty, and English language proficiency, must be considered.
- Teacher evaluation *should not discourage teachers from teaching less capable or motivated students*.
- Teacher evaluation systems should *encourage teachers to reflect on measures of student learning, interpreting results in relation to contextual factors, and use results to further improve student learning and motivation*.

- Teacher evaluation systems *should not encourage gaming of the procedures to produce false results*. Serious consequences tied to single, unreliable measures of student learning will encourage cheating and inappropriate manipulation of results.
- Teacher evaluation systems *need to include measures of all important learning goals*, not just those that are easily obtained. Often those measures that are efficient or most easily obtained are used while more complex or difficult measures that would assess other critical learning goals are minimized.

Comments and opinions from the teachers in this study support the validity of most of these principles. It was clear, for instance, that the teachers saw current evaluation efforts failing in large measure because of the use of assessments that did not connect well to what they have been responsible for. According to the teachers, there was considerable gaming of the current system, especially through the use of goal-setting and obtaining misleading percentages of students showing improvement. This was primarily a concern for using the pretest-posttest design over several weeks, whether or not these assessments were in the context of student goal-setting. These types of measures tend to be more standardized, and like large-scale end of year accountability tests, there was a disconnect between what more standardized measures of student learning show and targeted and actual learning that the teacher impacts. Teachers were also concerned about the accuracy of these types of measures since they were typically constructed by panels of local district teachers, or themselves, with little evidence of reliability or validity. As demonstrated in recent research, teachers do not trust interim test score data that are based on assessments that have flawed items or incomplete matching with learning standards (Abrams & McMillan, 2013).

While there is much research that demonstrates the influence of teaching on student achievement (Duncan & Spillane, 2008; Foorman, B. R., 2009; Nye, Hedges, & Konstantopoulos, 2004; Rockhoff, 2004) (indeed, many contend that teachers constitute the primary influence on learning), it seems clear that fair, valid, and accurate measures of student learning must include consideration of factors beyond the control of the teacher that result in low student motivation to learn or perform well on assessments of learning. This is a long-standing issue that is not easily addressed with the use of single sets of test scores that result in labeling and ranking of teachers. Evidence from these interviews, in fact, show that just the opposite will result from the use of a pretest-posttest type design of student learning and student goal-setting. The results from these measures alone will often not be accurate, resulting in mislabeling of effective teaching. To be fair, then, measures of student achievement must somehow be put in the context of these factors.

There is a well-known principle in educational testing that as the stakes rise, so does corruption of both intended and unintended consequences (Koretz, 2008). This principle appears to be manifest in current systems that include measures of student achievement such as interim or end-of-year tests, or goal-setting, to evaluate teachers. Some less than effective teachers, because of the high-stakes involved, appear to manipulate scores on these measures and goals to show competence. This consequence must be minimized if the use of measures of student achievement for teacher evaluation is to have a positive effect and result in the accurate labeling of ineffective teaching (one of the primary goals according to some). This is true for the use of classroom assessments as much as for more standardized measures, but perhaps more easily manipulated with pretests-posttests and accountability tests.

## Recommendations

The initial purpose of this research was to generate ideas about how classroom assessments could be used to provide evidence of student achievement for evaluating teachers. Because the use of classroom assessments for this purpose is not widespread, and used minimally with the teachers in the current study, it is not clear how these types of assessments would be influenced if used for teacher evaluation. At the very least, though, classroom assessments appear to meet some very important, established criteria for effective teacher evaluation. They provide multiple measures, show a clear connection between teaching and specific student learning, accommodate different types of classes and subjects, and allow for interpretations based on contextual factors that influence achievement.

Classroom assessments could be examined for coverage of standards and depth of learning, and feedback could be provided to teachers to help improve these dimensions of their assessments. This would require some type of peer review, which itself would be time intensive. An additional barrier could be the effect of such an approach on the relationship between students and teachers. Would teachers be less likely to be critical of student work and challenge students less because they would want the evidence to be positive, or would teachers be more challenging and motivated to provide more effective feedback to students so that they would show more growth in achievement? Would teachers set low goals with their classroom assessments to more easily show growth, purposefully write easy questions to obtain high scores, or grade papers more leniently? Perhaps the peer review procedure would mitigate these tendencies, but such consequences would need to be considered.

The use of student classroom achievement artifacts as evidence of student learning and effective teaching is not new, as illustrated most clearly by the use of these types of measures for evaluating National Board Certified Teachers. The NBPTS recognizes that compelling evidence of teaching is contained in the depth of learning that is illustrated in student achievement artifacts. Very simply, I believe students' learning is reflected most clearly in what is demonstrated on classroom assessments. Another more recent example is an effort in Tennessee (Tennessee Department of Education, 2013) that uses portfolios of student work for evaluating teachers in non-tests grades and subjects. In this model, teachers provide context, collect evidence from a purposeful sample of students and learning objectives, and upload evidence that is self-scored and reviewed by a peer. However, classroom assessments simply haven't been given much consideration for evaluating teachers.

The advantages of using classroom assessments suggest that some level of use of these types of assessments can be fruitful for teacher evaluation. Most significantly, classroom assessments, tests, quizzes, papers, and other artifacts that are already routinely used, have the ability to more validity show what teachers have accomplished with their students. What would be needed in a system that used classroom assessments to show credible evidence that would lead to accurate conclusions about teacher effectiveness, positive consequences on teaching and learning, and at the same time be efficient? Here is one possible scenario of a process, a series of steps, to consider for elementary teachers (secondary teachers could select one class each year):

1. Ask each teacher to identify, at the beginning of the year, one unit lesson taught each semester to be evaluated, each unit lasting approximately two to four weeks.

Teachers would need to verify by pledge that the units presented are generally representative of their teaching throughout the year.

2. By the end of the first week of class, ask teachers to provide a list of students rank ordered into three levels by aptitude and level of prior knowledge, with justification for the ranking, along with a summary of important contextual factors that could influence student achievement.
3. For each unit, the teacher provides a complete portfolio of all classroom assessments used during and at the end of the unit (not individual students).
4. Following completion of the unit, two students will be selected randomly from each of the three levels identified in Step 2.
5. The teacher will provide the specific classroom assessment results for all six students, with commentary on student motivation and contextual considerations for interpreting the results, and would provide overall judgments about the amount of student learning.
6. A panel of two trained elementary teachers, not in the same school, would review the assessments (e.g., questions, prompts), results for the six students, and teacher commentary to verify the match between content coverage and standards, the depth of student understanding that is demonstrated, and the reasonableness of teacher judgments about student learning based on results of the classroom assessments. Reviewers could also provide alternate interpretations of results and make suggestions to teachers for improving assessments, formative assessment, and instructional correctives.

7. Reviewer ratings would be provided to principals to use as one indicator of several that shows evidence of student learning that would be used toward the overall evaluation of the teacher.

The intent with this procedure is to provide an easily obtained, representative sample of student performance on classroom assessments, contained in a portfolio format that could provide a holistic presentation and evaluation of student learning. These assessments are already given; teachers would need to gather specific student performances at the end of each of the two units. Teachers would be allowed to provide commentary about the meaning of the results and contextual factors. The nature of the review of assessments and results would hopefully promote more effective use of classroom assessments by all teachers. By providing feedback to teachers on the depth of learning assessed, teachers would receive helpful information that could lead to a better alignment of classroom assessments and learning standards.

Classroom assessment, a practice that is already instituted, is directly aligned with teaching and contextually sensitive, offering a persuasive alternative to using value-added test scores, pretest/posttests, and student goal-setting. While further review and subsequent empirical research that can inform validity of using classroom assessments is needed, this type of evidence of student learning seems compelling when considered in light of generally agreed upon principles of effective teacher evaluation. Good classroom assessments are valid, contextually sensitive, multiple measures that could be efficiently gathered, mitigate teacher competitiveness, provide meaningful feedback from peers, and reflect what teachers are most responsible for. The literature about the use of large-scale test scores for teacher evaluation, supported by teacher voices such as those in this study,

shows that such measures are not effective, and may be harmful. More serious consideration of using classroom assessments is needed as an alternative that may provide more valid results with the potential of improving as well as documenting student learning.

## References


Abrams, L. M., & McMillan, J. H. (2013). Instructional influence of interim assessments: Voices from the field (pp. 103-130). In R. Lissitz (Ed.), *Informing the practice of teaching using formative and interim assessment*. Charlotte, NC: Information Age Publishing.

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. *Economic Policy Institute.* Retrieved from http://www.epi.org/publication/bp278/.

Berliner, D. C. (2014). Exogenous variables and value-added assessments: A fatal flaw. *Teachers College Record*, *116*(1), 1-31.

Chetty, R., Friedman, J.N., & Rockoff, J.E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. *National Bureau of Economic Research.* Retrieved from http://www.nbr.org/papers/w17699.

Danielson, C. (2010/2011). Evaluations that help teachers learn. *Educational Leadership, 68*(4), 35-39.

Darling-Hammond, L. (2013). *Getting teacher evaluation right: What really matters for effectiveness and improvement*. New York: Teachers College Press.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan, 93*(6), 8-15.

Duncan, G., & Spillane, J. (Eds.). (2008). Teacher quality: Broadening and deepening the debate. Evanston, IL: Multidisciplinary Program in the Education Sciences, Northwestern University, northwestern.edu/docs/teacher-quality.pdf.

Harris, D. (2011). *Value-added measures in education: What every educator needs to know*. Cambridge, MA: Harvard Education Press.

Foorman, B. R. (2009). Commentary: Informing teaching and learning policy. In Sykes, G, Schneider, B., & Plank, D. N. (Eds.). *Handbook of education policy research*. pp. 705-709. New York: Routledge.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *The Journal of Human Resources, 46*(3), 587-613.

Koretz, D. (2008). *Measuring Up: What Educational Testing Really Tells Us*. Cambridge, MA: Harvard University Press.

Lavigne, A. L. (2014). Exploring the intended and unintended consequences of high-stakes teacher evaluation on schools, teachers, and students. *Teachers College Record, 116*, 1-29.

Lavigne, A. L., & Good, T. L. (2014). *Teacher and student evaluation: Moving beyond the failure of school reform*. New York: Routledge.

Marshall, K. (2012). Fine-tuning teacher evaluation. *Educational Leadership, 70*(3), 50-53.

Marzano, R. J., Frontier, T., & Livingston, D. (2011). *Effective supervision: Supporting the art and science of teaching*. Alexandria, VA: ASCD.

Marzano, R. J., & Toth, M. (2013). *Teacher evaluation that makes a difference: A new model for teacher growth and student achievement.* Alexandria, VA: ASCD.

National Board for Professional Teaching Standards. (2011). *Student learning, student achievement: How do teachers measure up*? A report by the Student Learning, Student Achievement Task Force, Arlington, VA: National Board for Professional Teaching Standards.

Nye, B., Hedges, L., & Konstantopoulos, S. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*, 237-257.

Papay, J. P. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82*(1), 123-141.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data" *American Economic Review*, *94*(2): 247-252.

Steele, J. L., Hamilton, L. S., & Stecher, B. M. (2010). Incorporating student performance measures into teacher evaluation systems. Santa Monica, CA: RAND Corporation.

Stronge, J. H. (2010). *Effective teachers=student achievement: What the research says*. New York, Routledge.

Tennessee Department of Education (2013). Fine arts portfolio model: A new path to measuring growth in traditionally non-tested grades and subjects. Retrieved from http://team-tn.cloudapp.net/wp-content/uploads/2013/08/Fine-Arts-Portfolio-Model-Overview.pdf