NEE
Network for Educator Effectiveness
College of Education
University of Missouri

Using Student Achievement Data to Evaluate Teachers

Christi Bergin, PhD
April 23, 2015

The purpose of this white paper is to inform Network for Educator Effectiveness (NEE) members about current trends in the use of student achievement data in teacher evaluation. The decision about how to incorporate student achievement data into their teacher evaluation system is based on local policy. This document is intended to help NEE members understand their options better.

Background

The purpose of evaluating teaching effectiveness is to increase student learning. A common metric of student learning is achievement data (i.e., test scores). Thus, it is logical to use student achievement data to evaluate teaching effectiveness. The U.S. has had a short history of using achievement data to evaluate a school or district; the new movement is to use it to evaluate individual teachers' effectiveness. Indeed, the US Department of Education has used the Race to the Top funds and the ESEA waiver process to obligate states to use student achievement data as a "significant" part of teacher evaluation. We are beginning to hear discussion of using student achievement data with new teachers to evaluate their teacher preparation programs [1]. The hope is that student achievement data differentiates teachers better than old, inadequate evaluation systems that simply labeled teachers as "satisfactory" or not. These old systems have been justly criticized because almost all teachers received the same score which meant that great teaching went unacknowledged and poor teaching went unaddressed [2].

The basic idea is simple; students of more effective teachers learn more. However, quantifying student learning and then connecting the quantity of learning to specific teachers is far from simple. While there is consensus that student learning, in addition to quality of teacher practice, should be part of teacher evaluation, there is no consensus about how to include measures of student learning. Two common approaches for incorporating student achievement data into teacher evaluations are briefly summarized next – statistical approaches and student learning objectives.

Statistical Approaches: Student Growth and Value-Added Models

Student growth percentile models compare students' standardized test scores from one point in time to the next, relative to other academically similar students (or students with similar score histories). Value-added models (VAM) compare a student's predicted score with their actual score. The predicted, or probable, score is computed statistically from the student's prior test scores. In about 40% of states student demographics (e.g., poverty, home language, etc.) are statistically controlled in the models [3]. These two commonly used approaches are somewhat different, but both involve a statistical procedure for estimating a teacher's effect on students based on individual students' year-to-year growth in test scores. Individual student results are combined to create a teacher effect score. The idea behind both approaches is that teachers whose students grow less than comparable students or score below the predicted level are ineffective. Conversely, teachers whose students grow more than comparable students or who score above the predicted level are above average in effectiveness.

These models use large-scale state assessments [3], but may also use other standardized commercially available tests such as SAT 10 or ITBS. These alternative assessments tend to perform as well, if not better, than state assessments in these models. That is, they tend to correlate with the state assessments, they correlate modestly (about .50) with classroom observations, and they have low reliability or stability from

year to year (often less than .3) [4]. Some districts also use end-of-course exams (EOCs) in these models, but very little is known about the success of using EOC scores. One study, in the Pittsburg schools, found that EOC scores could distinguish among teachers better than standardized commercially available tests or state tests [4].

The important advantage of this approach is that results are considered objective, and comparable across teachers whose students take the same test. However, while these approaches may be adequate for evaluating large-scale programs, they are problematic for evaluating a single teacher. The problems include the following:

1. They do not inform growth.
    a. Even if they were fully reliable in distinguishing effective from ineffective teachers, they do not identify which teaching practices ineffective teachers need to improve[5]. In a study of 47 states, not one state representative was able to articulate how teachers could use the results to improve teaching practices [3].
    b. Analysis and feedback are delayed; results from spring testing are not typically given to teachers until the fall, when teachers have a new class[5].

2. Their reliability is positive, but meager. This means that an individual teacher's growth is likely to be somewhat similar from year to year, but there is considerable random variation [6, 7]. Results change depending on class, year, test and the statistical model used. The same teacher can have dramatically lower results if the number of English Language Learners, poor, or gifted students increases in his/her class [8]. Effectiveness will be over- or underestimated for a sizable number of teachers. For this and other reasons, their use has been cautioned in a public statement by the American Statistical Association [9].
    a. Reliability can be improved by using data across multiple years. This results in increased, but still somewhat low, reliability [4].
    b. Reliability can be improved by using a large number of students per teacher. This increases power to detect a small teacher effect amid random error. Having fewer students per teacher in the model's equation results in less reliability [7].

3. They are not applied to all teachers. Year-to-year testing using standardized assessments in most districts is available in only limited grades (typically 4-8) and subjects (typically math and reading). As a result, only 30% or fewer of teachers can be evaluated this way [3], which raises issues of fairness.

4. Their validity has been questioned for several reasons:
    a. The models do not use tests that are instructionally sensitive (i.e., aligned to what a teacher is teaching). Students' scores on global tests reflect many factors and life-time opportunity to learn, not just what an individual teacher taught in one class [10]. In addition, many tests do not measure valued skills, such as reasoning.
    b. Students are not randomly assigned to teachers. Some teachers are assigned academically talented students and others assigned struggling students in non-random ways [6, 11]. Teachers in high-poverty schools or classes tend to have lower value-added scores. To address this, statisticians may use "proportionality analyses," in which teachers in high poverty schools are only compared to other teachers in high poverty schools [12]. Statistically controlling for such pre-existing differences in students and schools helps level the playing field somewhat, but student achievement is still influenced by many factors that are not part of these statistical models [8, 11]. It is problematic to ascribe differences in test scores to an individual teacher. Correlational data does not prove causation.
    c. Students are often affected by multiple teachers in a particular subject each year, in addition to support specialists. This makes statistical parsing out of an individual teacher's

contribution difficult. In addition, it makes linking each student to a teacher challenging, error-prone, and time consuming. Districts need technical support and good data systems to attempt linking individual students with a "teacher of record." Some districts do this on a scheduled "claiming" day.

d. The models are not transparent. Their derivation is difficult for non-statisticians to understand [5, 13]. Many districts do not have personnel with the highly-specific skills to develop or implement them, so districts rely on universities or commercial entities to help them. Some of these entities keep their proprietary techniques secret.

Due to these problems, many experts have concluded student growth and value-added models are not appropriate as a primary measure of teacher effectiveness [e.g., 6, 10, 14]. Some argue that they should not comprise more than 20% of the evaluation of teachers and others argue that there is no evidence supporting their use at all. These approaches have been legally challenged [15]. Paige [16] argues that the legal cost of using them to make personnel decisions (e.g., raises or termination) outweigh any potential benefit.

There is consensus that these approaches need more investigation. In response, the Bill and Melinda Gates Foundation funded a large, multi-year study known as the *Measures of Effective Teaching* (MET) study. The MET study found that when over 1,500 teachers were randomly assigned to classrooms their value-added scores from the previous year predicted their value-added scores in the current year, supporting the utility of a value-added approach [17]. However, support was only found in math classes, but not language arts. In addition, the study has been criticized because randomization was faulty and because VAM was used in a low-stakes research context. It is not clear how it would function in a high-stakes evaluation context [18].

In response to the problems outlined above many districts are turning to SLOs.

Student Learning Objectives (SLOs)

In an SLO approach, teachers select assessments at the beginning of the year that are aligned with learning standards. Teachers establish a baseline on those assessments for their students, set learning targets, assess learning after instruction, and determine whether the learning targets were met. Up to this point, this process is what teachers should be doing routinely. (Many teachers will recognize these as "learning targets" and "SMART goals.")[1] The next step is new for most districts; teachers are then evaluated based on how many students achieve the target or whether a pre-specified goal is met (e.g., 85% of students reached the learning target.). Unlike VAMs, SLOs are not used in a statistical model.

Selection of assessments is a crucial part of the process. Teachers may select part of a state assessment, commercially available assessments, or teacher-generated tests, but they should have a core set of attributes. They should be (a) worthwhile, or target important learning objectives, (b) valid, reliable, and well-constructed, (c) include high-level thinking skills (i.e., high DOK[2] levels), and (d) be accurately scored [19]. As Popham [10, p. 71] points out, evaluators need to make sure that *valid inferences can be made about a teacher's ability to promote students learning* with the selected assessments.

Typically teachers are asked to set two SLOs (e.g., elementary teachers might have one for math and reading). They should be applicable to most students, but SLOs may be for a specific subgroup of students (e.g., English Language Learners). Special educators may collaborate with teachers in developing SLOs. SLOs may be for shorter or longer time frames (e.g., an instructional unit or a year-long course). Most

[1] SMART stands for specific, measureable, attainable, realistic, and timely goals. They are designed to guide the setting of objectives for instruction, as well as in work settings.

[2] DOK stands for depth of knowledge. It refers to a method of categorizing complexity of thinking, similar to Bloom's Taxonomy. Mere fact recall is at the lowest (1) level and cognitive functions like synthesizing and analyzing are at the highest (4) level.

districts encourage or even require teams, rather than individual teachers, to develop SLOs. Some even evaluate grade-level teams, not individuals, on meeting SLOs [20].

To evaluate teachers using SLOs, administrators apply a scoring rubric (which may be provided by the district, or mutually agreed upon with the teacher), and assign a rating. For example, the scoring rubric may be on a scale of 1 to 4 with the high anchor being "1.5 years of growth are achieved," or "90% of content mastery is achieved," or "80% of students move from level X to level Y." Some rubrics rely on more judgment than specific numbers (e.g., exceptional number of students make outstanding progress vs. goal not met and little student progress made). Depending on the priority of the school, one SLO may be more heavily weighted than a second SLO (e.g., 60% vs. 40%). Ideally, the SLO scoring process leads to a conversation between administrator and teacher regarding how to improve teaching practices.

SLOs have important advantages. They can be used in any grade and subject. If carefully selected, assessments are instructionally sensitive. They provide opportunity to measure abilities, like critical thinking skills or artistic creativity, not typically measured by state assessments. SLOs give teachers ownership of instructional goals, encourage data-driven instructional decisions, and promote collaboration over instructional goals between teachers, grade-level teams, and principals. Some, but not all studies, find that teachers report SLOs improve instruction by focusing on student needs [4]. Thus, they serve both instruction and evaluation. However, these dual purposes can be conflicting. Setting low SLOs can boost evaluations, but not are not ideal for instructional planning.

Despite these advantages, there are several problems with the use of SLOs:

1. They add burden to teachers who must make informed selection of assessments and monitor data collection. Many teachers and principals do not have adequate assessment literacy. Teachers need training on selecting SLOs. Teachers need easy-to-use software that links student information to teacher evaluation with prior years' data [4, 10].

2. Teachers and principals may game the system by selecting low targets to get higher evaluations. Ideally, SLOs are ambitious but attainable. Yet, teachers complain that others choose lower targets, and, partly because of this, view SLOs as the least effective form of evaluation [4]. To address this problem many districts require SLOs to be approved by the principal and district personnel, or an SLO "rigor" oversight team [21].

3. Important comparisons cannot be made if SLOs are not consistent across classes.

4. SLOs may not be adequately discriminating; 55-95% of teachers meet their targets. The longer a district uses SLOs, the more teachers meet their targets. Thus, they may not be very useful for identifying effective teachers from those needing support [4]

5. Currently little is known about their reliability or validity when used in evaluation. They are not yet adequately developed for use in high-stakes decisions. Some districts roll them out in a low-stakes environment before using them for high-stakes evaluation [4].

One of the key issues with use of SLOs is how to balance standardization with individualization and ensure rigor. Following are some examples of different approaches [4]:

- New York: The state provides a list of approved SLOs.
- Indiana: The state provides optional SLOs. Teachers set two SLOs for one class in their first year of implementation; but for all classes thereafter.
- Georgia: The district creates two or three SLOs for courses, not individual teachers. The state approves the SLOs.
- Washington DC: The district provides suggested assessments for each grade and subject.

- Ohio: Teams of teachers develop SLOs; SLO evaluators approve them.
- Austin: Teachers develop SLOs. The principal approves them and a district team assesses them for rigor.
- Denver: Teachers develop two SLOs each year. The principal approves them and submits them to an online forum.

Incorporating Student Achievement Data into Evaluations

States and districts may use either, or both, of these two basic approaches. In some states, teachers in "tested" subjects or grades are evaluated with a combination of VAM and SLOs, whereas teachers in other subjects or grades are evaluated only on SLOs. Clearly, both approaches have advantages and disadvantages. This is true of other types of measures as well (e.g., classroom observations). For this reason, multiple measures should be used to evaluate teachers [10]. Indeed, the American Educational Research Association has taken the position that high-stakes decisions should not be made on the basis of test scores alone. Multiple sources of information should be considered to enhance the overall validity of such decisions [22].

When multiple sources of evidence about teaching effectiveness are gathered, states or districts must determine how to combine them. There is no clear consensus among experts on how to incorporate and weight evidence of student learning into a larger summative evaluation (or even whether to include it at all). States and districts vary enormously in the extent to which student achievement data is weighted in the evaluation process, from 10% to 50% of the evaluation formula. Another NEE white paper addresses the issue of weighting.

Resources

- For a comparison of VAM and other statistical approaches to analyzing student achievement growth see the CCSSO's practice guide [23].
- For a video explaining the basic concepts of the VAM by the University of Wisconsin-Madison see https://www.youtube.com/watch?v=UaGuaSw7OQI.
- For examples of how two districts address SLO use, see Denver and Charlotte-Mecklenburg. As the first district to document use of SLOs, Denver is often used as a model. Their website has guidelines for teachers to develop SLOs and a rubric for assessing them at sgo.dpsk12.org. See also an SLO Guide at http://www.ctacusa.com/PDFs/CMSSLOGuide-2008.pdf.
- For an accessible discussion of these issues and teacher evaluation overall see Pophams' book, *Evaluating America's teachers: Mission possible?*, in the reference list.

References

1. Plecki, M.L., A.M. Elfers, and Y. Nakamura, *Using Evidence for Teacher Education Program Improvement and Accountability: An Illustrative Case of the Role of Value-Added Measures.* Journal of Teacher Education, 2012. **63**(5): p. 318-334.
2. Weisberg, D., et al., *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*, 2009, The New Teacher Project: Madison, WI.
3. Collins, C. and A. Amrein-Beardsley, *Putting growth and value-added models on the map: A national overview.* Teachers College Record, 2013. **116**(1): p. 1-32.
4. Gill, B., J. Bruch, and K. Booker, *Using alternative student growth measures for evaluating teacher performance: What the literature says*, 2013, Mathematica Policy Research.
5. Goldring, E., et al., *Make room value added: Principals' human capital decisions and the emergence of teacher observation data.* Educational Researcher, 2015. **44**(2): p. 96-104.
6. Darling-Hammond, L., et al., *Evaluating teacher evaluation.* Phi Delta Kappan, 2012. **93**(6): p. 8-15.

7. Ballou, D. and M. Springer, *Using student test scores to measure teacher performance: Some problems in the design and implementation of evaluation systems.* Educational Researcher, 2015. **44**(2): p. 77-86.
8. Berliner, D., *Exogenous variables and value-added assessments: A fatal flaw.* Teachers College Record, 2014. **116**(1): p. 1-31.
9. ASA, *ASA Statement on Using Value-Added Models for Educational Assessment*, April 8, 2014, American Statistical Association.
10. Popham, W.J., *Evaluating America's teachers: Mission possible?*2013, Thousand Oaks, CA: Corwin.
11. Konstantopoulos, S., *Teacher effects, value-added models, and accountability* Teachers College Record, 2014. **116**.
12. Koedel, C., *Proportionality in district, school, and teacher evaluation systems*, November 2014, Institute of Public Policy: University of Missouri.
13. Gabriel, R. and J.N. Lester, *Sentinels guarding the grail: Value-added measurement and the quest for education reform.* education policy analysis archives, 2013. **21**: p. 9.
14. McCaffrey, D.F., et al., *Evaluating value-added models for teacher acountability*, 2003, RAND Corporation: Santa Monica, CA.
15. Strauss, V., *High-achieving teacher sues state over evaluation labeling her 'ineffective'*, in *Washington Post*Oct 31 2014: Washington.
16. Paige, M.A., *A legal argument against the use of VAMs in teacher evaluation.* Teachers College Record, 2014.
17. Kane, T.J., et al., *Have we identified effective teachers? Validating measures of effective teaching using random assignment*, 2013, Bill & Melinda Gates Foundation.
18. Rothstein, J. and W.J. Mathis, *Review of "Have We Identified Effective Teachers?" and "A Composite Estimator of Effective Teaching: Culminating Findings from the Measures of Effective Teaching Project"*. National Education Policy Center, 2013.
19. Herman, J.L., M. Heritage, and P. Goldschmidt, *Developing and Selecting Assessments of Student Growth for Use in Teacher Evaluation Systems.* Assessment and Accountability Comprehensive Center, 2011.
20. Lacireno-Paquet, N., C. Morgan, and D. Mello, *How States Use Student Learning Objectives in Teacher Evaluation Systems: A Review of State Websites.* Washington, DC: US Department of Education, Institute of Education Sciences, 2014.
21. Gill, B., et al., *Alternative student growth measures for teacher evaluation: Profiles of early-adopting districts*, 2014, Mathematica Policy Research.
22. Joint Committee of AERA, APA, and NCME, *Standards for Educational and Psychological Testing*2014, Washington DC: AERA.
23. Castellano, K.E. and A.D. Ho, *A practitioner's guide to growth models*, 2013, Council of Chief State School Officers.