

# Considerations for Using Assessment Data to Inform Determinations of Teacher Effectiveness

Chris Domaleski and Richard Hill, Center for Assessment<sup>1</sup>

April 29, 2010

## Introduction

Recently, the United States Department of Education (USED) released a notice inviting applications (NIA) from consortia of states to develop comprehensive assessment systems (hereafter *common assessments*). Among other requirements, the NIA specifies that a common assessment must produce data to inform determinations of principal and teacher effectiveness. This white paper is written to assist consortium members in indentifying and addressing the conditions, challenges, and opportunities to help them respond to this requirement.

While there is considerable overlap between issues related to teacher and principal effectiveness, we primarily limit our focus in this paper to issues related to teacher effectiveness. In so doing, we note that those issues associated with principal evaluation are also associated with teacher evaluation. However, teacher evaluation presents challenges above and beyond those associated with principal (school-level) evaluation. If the focus was solely at the school-level, some of the issues presented herein would not arise.

Finally, it is important to acknowledge that the NIA identifies two clear purposes for using assessment data to inform determinations of effectiveness. The first is to support evaluation, meaning that the results should be used as a component in the appraisal of professional performance. The second purpose is to inform professional development and support. That is, the data should be useful to help identify and encourage practices that are most promising and to discourage those that are least effective.

## Design Considerations

Before examining the use of the data, we begin by addressing the elements of the of the assessment system that must be in place in order to produce results that have the potential to inform teacher effectiveness.

### Assessment Characteristics

A starting point to support the claim that higher test scores reflect more effective instruction, is a test that is both sensitive to effective instruction<sup>1</sup> and represents valued outcomes. That is, given a fixed starting position, test scores should be higher for students who have received effective instruction on the desired outcomes than it will be for students who either have received ineffective instruction on the

---

<sup>1</sup> Popham (2007) has proposed a framework for evaluating instructional sensitivity of assessments.

desired outcomes or who have received instruction (effective or not) on outcomes other than the desired ones.

Each element is necessary but insufficient. An assessment measuring the 'right' outcomes but is not instructionally sensitive, fails to yield useful information about teacher effectiveness. On the other hand, if the assessment does not represent those standards that are most important for teachers to teach and students to learn, it may incentivize the wrong teaching practices.

If the assessment is to be a trustworthy measure of valued knowledge and skills, it should have sufficient breadth to cover the full range of content and sufficient depth to address these standards beyond a superficial level. It is unlikely that this can be accomplished with an assessment comprised of multiple-choice items alone. The inclusion of constructed response items and performance tasks will better enable the assessment to more completely address the construct of interest.

Beyond the content represented on the assessment, the range of performance measures produced must be sufficient. With traditional criterion-referenced tests it is common to construct the assessment such that most of the information is placed around the cut scores separating one or more performance levels. This is useful when the primary objective is to maximize the accuracy of performance level classifications. However, such a test is often ill-suited to determine degrees of performance within levels, particularly for the highest and lowest levels. If the assessment is to produce useful information about student progress to inform educator effectiveness, it must have 'high ceiling' and a 'low floor' to measure performance across a broad performance range. If the range is not sufficiently broad, the assessment will not reliably detect gains between multiple assessments for students of high or low ability.

### **Measuring Student Progress**

It is generally acknowledged that any use of test data to inform teacher effectiveness should control for prior performance. Therefore, the assessment system must produce a measure that reflects the progress or growth of the student during the period of time the teacher provided instruction. Broadly, there are two primary elements that must be in place to accomplish this goal: 1) availability of one or more prior scores and 2) application of a suitable analytic method.

To start, the structure of the assessment system should be such that one or more suitable prior scores are available. One way to accomplish this is to use a score from the end of the previous year. This is a fairly straightforward approach for an assessment system based on coherent end-of-year assessments. If there are assessments at the end of each of grades 3-8 it may be possible to use the previous year's score as a baseline for determining progress starting in grade 4. However, this approach is often more complicated for high schools where assessment systems may be structured around either a set of comprehensive assessments given at a single point (e.g. end of grade 11) or a set of end of course assessments (EOC) administered variably as a student encounters the course. With both approaches the timing presents an issue. For example, it would be very challenging to evaluate any one teacher based on performance changes during the three year gap between a grade 8 test and a grade 11 test. Moreover, the variability in course sequence and the lack of coherence may render the EOC approach

prohibitive, unless a pre-test is administered in these courses. The sequence problem simply describes the fact that students are often permitted to take courses at different grades and in a different order (e.g. one student takes algebra in grade 8 and geometry in grade 9; another student takes geometry in grade 9 and algebra in grade 10). The coherence issue means that it may not be reasonable to assume that two EOCs share the same construct such that one is meaningfully related to the other.

Another approach is to use interim assessments administered at the beginning of and/or at multiple times during the term of instruction. This method may better control for extraneous influences, such as the effect of student gains or losses between terms. However, interim assessments are often taken following some amount of instruction. The shorter the timeline between the pre and post test, the less true variance there is to account for, which will diminish reliability and lessen the effectiveness of this approach.

In any case, if this approach is selected, is important to ensure that the interim assessment or pre-test is well suited for its intended use. Such a test should be well-correlated with the outcome assessment. Additionally, to the extent it represents the construct of interest, claims that gains are associated with instruction are better supported.

The second consideration for producing a meaningful growth score is the implementation of an appropriate analytic method. There are a variety of approaches to consider and a full treatment of this topic is beyond the scope of this paper. However, we identify three general methods to introduce options. The most straightforward approach may be to produce a gain score. This involves simply computing the difference between the pre-test and the post test. Such an approach relies on assessments that share a developmental or vertical scale. While gain scores may be intuitively appealing, many researchers remain skeptical that vertical scales can support such interpretations and many do not consider this method a promising approach. Another analytic approach that is often associated with teacher effectiveness is a value-added model (VAM)<sup>2</sup>. VAMs are a family of regression based analytic techniques in which certain variables are included in the model in an attempt to account for unrelated variance to better isolate the effect of interest – in this case, a teacher’s contribution to achievement. Still another option is to compute a normative measure of growth – the Student Growth Percentile (SGP) (Betebenner, 2009). The SGP method examines performance of students with identical prior achievement scores and computes a percentile for each student indicating the probability of that outcome given the student’s starting point (Betebenner, 2009). This can be used to gauge whether or not the student’s growth was atypically high or low.

## **Challenges**

Even with a well-designed assessment system that produces a trustworthy measure of student progress, a number of challenges must be addressed in order to move to the next step of associating those results with teacher effectiveness. Those challenges may be grouped into two categories: attribution and utility.

---

<sup>2</sup> For more information regarding use of VAMs for teacher effectiveness, the reader is referred to Braun, Chudowsky, & Koenig, 2010 and McCaffrey et al. (2003).

## Attribution

Attribution refers to the challenge of linking teacher behavior to student outcomes. Impediments to attribution include:

- *Limited involvement of grades and subjects.* If a teacher has no (or a limited number of) students being tested, results cannot be produced. Unless the scope of the assessment system is expanded well beyond what is required in the NIA, many teachers would be excluded from the model such as those teaching in grades K-2 and those teaching subjects not tested such as social studies and science. At the secondary level, it is conceivable that a relatively small proportion of teachers will be included. In other cases, results will only be available for a portion of the material taught, such as for a middle school teacher who provides instruction in mathematics and science. Similarly, if principals are included, it is important to recognize that the data used for evaluation would be derived from a subset of the areas over which their leadership extends.
- *Assigning accountability.* An additional challenge is figuring out which teacher should be held accountable for a student's performance. This involves establishing a data system that connects test scores to the instructor (as opposed to, for example, the homeroom teacher). Moreover, it is not uncommon for students to receive instruction over tested material from multiple teachers, such as with a student who receives additional support services outside the traditional class. Some researchers have attempted to fine-tune the attribution problem by assigning proportional responsibility for students to all or most individual educators in the school. However, this seems like an accounting scheme to try to avoid (not solve) the attribution challenge. In the end, significant issues, such as how to handle students who move multiple times during the year or whether/how much to 'weight' responsibility for students who spend a portion of the term in multiple schools or classrooms needs to be addressed.
- *Extraneous factors.* It is not enough simply to link scores to educators, it is critical to establish a causal link. That is, how can we know that it was the teacher's behavior that led to the observed gains and not other factors? These factors might include those that advantage performance, such as having a strong teacher the previous year or availability of home enrichment, or factors that mitigate performance, such as a student who infrequently attends class or experiences a family crisis during the instructional term. Establishing causal attribution is a non-trivial task and typically involves engaging in systematic data collection and research to both strengthen the association between the hypothesized antecedent (i.e. quality instruction) and the consequent (i.e. increased test scores) and to rule out rival explanations for the outcome.

## Utility

Some of the challenges in this category overlap with those related to attribution, but they are presented separately to describe specific threats to forming meaningful, actionable interpretations from the

results. In other words, are the results credible and can they help teachers improve student achievement? Threats to utility include:

- *Lack of random assignment.* Growth scores assume that students have been randomly assigned to teachers and that the probability of a student's score at the end of the year is predictable, given the background information about the student. However, it is a common and constructive practice to assign students to a particular teacher specifically because they are more likely to have success with that teacher.
- *Sampling error.* Student-level growth scores have considerable variability associated with them compared to status scores. Moreover, sampling error is directly related to the number of observations - as the sample size increases, the variability reduces. This problem is somewhat assuaged when computing a growth score for a school across several teachers and grades. However, at the classroom level, where outcomes are likely based on 10-30 students, scores may vary substantially based on sample characteristics alone, as opposed to 'true' gains.
- *"What is" versus "What should be."* Even in the best case, growth interpretations reveal the *relative* effectiveness of teachers (compared to other teachers), which is not necessarily a meaningful finding. That is, we might conclude that a student with 'above average' growth performed better than what is typically observed or what one would predict based on students with similar prior scores. Importantly, they do not describe how much students could be or should be growing with effective instruction.

This issue interacts with the lack of random assignment problem. If, for example, economically disadvantaged (ED) students were disproportionately assigned to less effective teachers, their expectations for future growth will be less than what it would have been if the students had more effective teachers. This could lead to at least two problems. First, it is not correct to assume that teachers with equivalent VAM scores were equally effective (the teacher of the ED students who achieved an 'average' score is likely less effective than another teacher of non-ED students with similar results). Second, it leads to the prediction of lower gains for ED students, which could reinforce an unacceptable status quo.

- *Summary scores mask important differences.* Just as two classes with the same mean scale score may have dissimilar distributions, classes with the same mean gain score may be quite different. Consider, for example, three classes that have a mean VAM score of 10. In one class each student's gain was 10. In another, 10 ED students did not gain but 10 non-ED students gained by 20 points. In the third, a single student experienced an extraordinarily large gain of 200 points, but no other students improved. The point is, while all of these classes share the same overall result, a strong case can be made that each result is not equally valued.
- *Difficulty identifying effective practices.* If results indicate that a teacher was more or less effective than other teachers, the next logical question is, "what specific practices contributed

to this?” It is doubtful that a teacher is strong or weak in all areas and it is important to isolate the actions that had the most desired impact in order to capitalize on gains or mitigate/reverse losses.

## **Recommendations**

A well-known expression cautions one to avoid letting perfection be the enemy of good. Regardless of the caveats addressed in this paper, the NIA provides an unprecedented opportunity to make progress in developing solutions. It may be possible to design an approach that, if not perfect, improves on current teacher evaluation practices and provides feedback that is more useful for professional development.

In that context we offer the following recommendations to guide responses to the NIA:

- *Specify claims in the Theory of Action.* Include teacher evaluation and support in the overall theory of action (TOA) that guides the assessment and accountability system. It is important to specify the overarching and supporting claims as well as the mechanisms hypothesized to achieve desired outcomes (Marion, 2010). A well specified and credible TOA will aid in designing an approach that is purposeful and amenable to evaluation.
- *Develop a robust research plan.* The claims presented in the TOA should be falsifiable and accompanied by a robust research plan to evaluate and refine the assumptions. For example, if we claim that the most effective teachers have students who exhibit the largest gains, do we identify the same teachers as effective from year to year? Does the claim hold up for classes comprised of students with different characteristics (e.g. gifted students as well as persistently low performers)? Are instructional approaches similar for teachers rated equally effective? What additional sources evidence (e.g. external test scores) support this conclusion? The validity of such claims increases as more sources of evidence point to the same conclusion.
- *Establish credible performance expectations.* As discussed, equivalent outcome measures may not indicate equally effective instruction and summary results can mask important differences. Before establishing standards for ‘good enough’ performance develop a plan to investigate the amount of growth that can be reasonably expected under various conditions and explore ways to go beyond summary results to identify and assign value to the outcomes that are most desirable.
- *Include multiple sources of information.* While assessment results may be one source of information to help determine teacher effectiveness, it should not be the only component. Nothing in the NIA suggests that determinations cannot be or should not be determined by multiple indicators such as supervisor or peer reviews, observations, attainment of valued credentials etc.

It may be possible to draw on existing research to develop methods that incorporate multiple indicators (see, for example, Danielson & McGreal, 2000). Additionally, consortia may be able to describe a process for investigating and evaluating alternatives to develop a solution.

- *Quantify and acknowledge sources of error.* As addressed previously, there are both measurement error and sampling error associated with growth scores at the class and school level. Before one can determine whether or not the error is prohibitively large to support intended uses, it is important to quantify and explicitly report the error. Hill and DePascale (2002) propose some methods to evaluate the reliability of school level accountability decisions that may be applicable in this context.
- *Recognize that higher stakes creates greater burden.* To the extent that a consortium, SEA, or LEA proposes to use results for high stakes applications (e.g. merit pay, grounds for termination etc.) the burden to demonstrate that the system is fair and accurate is increased. Moreover, as the stakes elevate, it is important to guard against unintended consequences such as promoting competition among teachers instead of cooperation or assigning 'poor' teachers to grades/subjects not assessed.

## References

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Assessment: Issues and Practices*, 48 (4), pp. 42-51.

Braun, H., Chudowsky, N., and Koenig, J. A., editors (2010). *Getting Value Out of Value-Added: Report of a Workshop*. Washington, D.C.: The National Academies Press. Available at <http://www.nap.edu/catalog/12820.html>.

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation: To enhance professional practice*. Alexandria, VA: Association for Supervision and Curriculum Development

Hill, R., & DePascale, C. (2002). *Determining the reliability of school scores*. Paper commissioned by the Council of Chief State School Officers.

Marion, S. (2010). *Developing a Theory of Action: A Foundation of the NIA Response*.

McCaffrey, D. F., J. R. Lockwood, D. M. Koretz, and Laura S. Hamilton. (2003). *Evaluating Value Added Models for Teacher Accountability*. MG-158-EDU. Santa Monica, CA: RAND.

Popham, W. J. (2007) Instructional sensitivity: Educational accountability's dire deficit. *Phi Delta Kappan*, 89 (2), 149–155.

---

<sup>i</sup> This paper is one of a series of brief documents produced by the Center for Assessment designed to assist state consortium leaders in thinking about and preparing their responses to the United States Department of Education Notice Inviting Applications for Comprehensive and High School Assessment System grants. Funding has been provided by the Bill & Melinda Gates Foundation. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Gates Foundation.