CECR

Center for
Educator Compensation
Reform

Resolving Some Issues in Using Value-Added Measures of Productivity for School and Teacher Incentives

Ideas from Technical Assistance and TIF Grantee Experience

The Harvesting Project

Anthony Milanowski
Center for Educator Compensation Reform
Wisconsin Center for Education Research
University of Wisconsin-Madison

The input of Mike Christian of the Value-added Research Center at the University of Wisconsin-Madison is gratefully acknowledged.

Resolving Some Issues in Using Value-Added Measures of Productivity for School and Teacher Incentives

Ideas from Technical Assistance and TIF Grantee Experience

The U.S. Department of Education through the Center for Educator Compensation Reform (CECR) supported the work described herein. The opinions expressed are those of the author and do not necessarily reflect the view of the U.S. Department of Education, CECR, or the institutional partners of the Center.

Introduction

Although many researchers and policy analysts (e.g., Harris, Glazerman et al., 2011; 2010) consider value-added to be the state of the art in school and teacher productivity measurement, only a minority of TIF Round 1 and 2 grantees used value-added as a measure of school or teacher performance. Fourteen of the 34 grantees proposed to use school-level value-added and were using this for the 2009-10 school year. Thirteen proposed to use classroom value-added, but in the 2009-10 school year, only 10 did so (see table 1).[1]

Table 1. Use of Value-Added[2] by Round 1 and 2 TIF Grantees

	Originally Proposed	Used as of 2009-10 School Year
School-level value-added	14 of 34	14 of 34
Classroom-level value-added	13 of 34	10 of 34

Why did more grantees not use value added? We found in our initial harvesting work that one of the main reasons for this was lack of appropriate administrative databases with links between teachers and students (Milanowski, Witham, Schuermann, Kimball, & Pietryka, 2010). Watson, Witham, and St. Louis (2010) discuss this issue in the companion Harvesting Project paper (2010). But there were other difficulties as well, including limited grantee capacity to develop and run complex value-added models, low buy-in from some stakeholders, schools or districts too small to develop meaningful value-added estimates, perceived lack of compatibility with state Adequate Yearly Progress (AYP) models or accountability systems, and simply a lack of comfort with the complexity of the technology needed to operate a performance-based compensation system (PBCS).

1 See the appendix for information on which grantees use value-added, as indicated by original proposals and grantee self-evaluation reports.

2 Note that we are not counting two kinds of growth measures as value-added models: (1) measures based on changes in a student's level of proficiency over time and (2) models that rely on teachers and/or principals to set goals for test score improvement based on post- and pretests with similar scales.

Some of the grantees who had planned to use classroom value-added found that they were not immediately able to do so. Value-added also requires more expertise to implement than many grantees initially possessed. Most of the grantees using value-added (11 of the 14) partnered with a vendor or consultant to produce the actual value-added estimates. In the first two rounds of grants, SAS in Schools was the most commonly used vendor, working with eight grantees. Mathematica Policy Research did value-added for three grantees; RAND worked with Pittsburgh; and the Value Added Research Center at the University of Wisconsin-Madison (VARC) provided value-added estimates to Chicago and will be providing them for Hillsborough, FL. Three grantees have developed and run their own models (Amphitheater, AZ, Charlotte-Mecklenburg, NC, and Dallas, TX).

In order to assist those who might be thinking about using value-added productivity measures as part of a PBCS, this harvesting paper discusses four barriers TIF 1 and 2 grantees have encountered and how they were addressed:

1. Concerns about the "right" value-added model
2. Uncertainty of value-added estimates for smaller schools and classrooms
3. Applying value-added when there are very few schools or classrooms to compare
4. Setting appropriate cutoffs for bonus payments

The information in this paper uses as a basis CECR experiences in providing technical assistance to grantees, work done by VARC researchers, and the rapidly developing literature about value-added. We do not intend this paper to be a technical presentation about value-added methodology, though we assume some familiarity with value-added concepts and use and cite some technical studies. Technical papers that vendors or consultants have written were an important data source for this paper.

Note also that this paper does not recommend any particular vendor or specific value-added model.

A. What Is the "Right" Value-Added Model?

Value-added estimates are superior to the use of student attainment or the cross-cohort comparison of proficiency levels used to determine AYP under the *No Child Left Behind Act (NCLB)* (Meyer 1992). The value-added method's major contribution is that it controls for pre-existing differences in achievement that are outside of the control of the teacher or school and, directly or indirectly, for differences in socio-economic status (SES) that influence achievement. When used with tests that are vertically scaled or equated, value-added can reflect growth in learning that represents the increment of learning that a student achieved during the school year, given prior knowledge and other factors that were already present. However, there are several versions of value-added models and no consensus yet among users or experts as to which is the best.

Several TIF grantees have expressed reluctance to using value-added measures of productivity in educator incentive systems because there is no consensus in the research community on the "right" value-added model. This concern is understandable. A recent National Research Council report (*Getting Value Out of Value-added: Report of a Workshop*; Braun, Chudowsky, & Koenig, 2010) recognized this lack of consensus. However, some consensus is developing on a few of the modeling issues, as indicated by the work done by and for TIF grantees. This section discusses four issues: (A) the inclusion of student demographic characteristics in the model, (B) pretest measurement error, (C) how to account for student mobility, (D) and how to handle mid-year testing.

A. *Should Student Demographic Characteristics Be Included in the Model?*

The value-added systems that TIF grantees use vary as to whether and which student demographic factors (e.g., sex, ethnicity, free or reduced-price lunch eligibility, special education participation) serve as controls in the model. The EVAAS model used by Houston and many of the grantees using the TAP model do not include demographic variables. The models used by Dallas; Chicago; Washington, DC; Memphis; the EPIC Charter school consortium; Amphitheater; and the new South Dakota model all include demographic variables. The designers of the Dallas model have historically gone to considerable lengths to control for demographic factors, and the model includes them at both the student and school levels.[3]

Whether student demographics should be included in a value-added model is controversial. Proponents of the EVAAS model (e.g., Sanders, Wright, Rivers, & Leandro, 2009) generally argue that this control is unnecessary, due to the use of multiple years of student test data in the EVAAS model.[4] They also argue that including student demographic information lowers expectations for students from groups that have historically shown lower achievement. Meyer and colleagues (Meyer & Christian, 2008; Meyer & Dokumaci, 2010) counter by arguing that when schools or classrooms have substantially different demographics, adding these controls makes for a fairer comparison and allows an explicit estimate of how much better or worse districts or schools are doing in facilitating learning for these students. Knowing how much of an effect demographic characteristics have on student learning helps focus attention on differential achievement and allows measurement of the success of interventions aimed at reducing the achievement gap. Deciding to include demographic variables is probably as much a policy decision as a technical decision. For example, in Dallas, controlling for student demographics has been important to educator acceptance of value-added as a performance measure. In contrast, Hillsborough County, Florida, does not do so, nor does Tennessee in its statewide value-added system. Tennessee provides schools and teachers with tools to compare value-added estimates for their students by poverty level, ethnic heritage, and other characteristics, to identify problematic patterns of growth.

Adding these controls may make relatively little difference in the overall distribution of productivity estimates if the distribution of these characteristics is similar across schools or classrooms. Mathematica's report on Memphis (Potamites, Chaplin, Isenberg, & Booker, 2009) found that the correlation between school rankings with and without student demographics was high (0.996). Similarly, their report on value-added estimation for the EPIC school consortium found a correlation of 0.988 for schools (Potamites, Booker, Chaplin, & Isenberg, 2009; Potamites, Chaplin, et al., 2009). Mathematica's findings are consistent with some prior research (Ballou, 2005; Ballou, Sanders, & Wright, 2004; Kane & Staiger, 2008). However, the Mathematica researchers also found that adding controls for student demographics increased the precision of elementary school value-added estimates, though this effect was smaller than adding a year of student data. Using a model much like Mathematica's,

3 In the Dallas model, school averages are used to represent the possible effects of school context (e.g., having a high proportion of poor students) on achievement, over and above any effect of individual student characteristics (such as poverty). See Bembry, Weerasinghe, & Mendro (1997) or Webster, Mendro, Orsak, & Weerasinghe (1998).

4 Note that Sanders (2006) reports that in one district the EVAAS classroom value-added estimates had negligible correlations with SES indicators, compared to models using one year of data. However, in another district, substantial correlations remained, though they were considerably smaller than those with estimates from a one year of data model. Some of this difference could be due to the homogeneity of the district. If student SES is relatively similar across schools or classrooms, adjusting for SES is likely to have less effect.

Kane and Staiger also found that including student demographics increased the precision of teacher effect estimates. Work by the VARC (e.g., Meyer & Dokumaci, 2009) found that, in models estimating one year of value-added, student demographic characteristics related significantly to test scores even after controlling for pretest scores. In that case, including or not including controls is likely to affect where individual teachers fall in the performance distribution. This will most likely happen when test scores are correlated with student characteristics such as poverty, and when student characteristics differ substantially across classrooms or schools. For example, if high-poverty students tend to learn less in each grade, classrooms with more poor children will have lower post-tests for any given pretest, and teachers will have lower value-added. Controlling for poverty reduces the expected post-test score and raises the estimated value-added.

It appears that adding student demographic characteristics as controls would be useful in models of value-added that attempt to measure teacher or school productivity during a single school year and do not include multiple years of student test data. Schools and districts that include these controls may not only increase perceived fairness among educators, but will also likely improve the precision of classroom or school effect. This is important because improving precision makes it easier to distinguish performance differences among schools and classrooms. Given the possibility of increased precision and the concerns educators have about fairness, it is probably better to over-control for factors outside their immediate influence. Schools and districts could address concerns about lower expectations, as Meyer has suggested, by reporting and tracking the size of significant negative effects of ethnic, sex, or poverty factors.

B. Pretest Measurement Error

Users must also consider how to deal with measurement error in the pretest. Since the pretest is almost always the most important predictor of post-test performance, users should not underestimate its importance. The more error in the pretest, the weaker the relationship between pre- and post-test appears. This leads to an underestimate of the effect of prior achievement, so that grantees erroneously attribute more variation in achievement to other sources, including the school or teacher. Hanushek and Rivkin (2010) reviewed six studies of classroom or teacher-level value-added and found that test measurement error is nearly as great a cause of variation as true variance in effectiveness in models using two years of test data (post- and pretest), suggesting that test measurement error is a serious issue.

Value-added models used by TIF grantees address pretest error in three different ways. The Mathematica models for Memphis and the EPIC school consortium used an approach called instrumental variables that uses the pretest for a different subject to estimate the part of the original pretest that does not have measurement error. They found that using this method increased the size of the coefficient for prior test score from 0.57 to 0.88 in the Memphis data (Potamites, Chaplin, et al., 2009).[5]

The models used by Amphitheater and Chicago control for measurement error by adjusting the coefficient on prior achievement upward by a factor representing the size of the average measurement error for the test population. The test vendor provides the average measurement error, standard

5 More recently, Mathematica has used a similar pretest measurement error correction model to that used in Chicago and Amphitheater in Washington, DC. See Isenberg & Hock, 2011.

error of measurement, or test reliability, which can be used as the basis for adjustment.

The EVAAS approach to pretest measurement error is to add additional years of test scores to the model. Sanders (2006) reported that, based on real data and simulation results, a minimum of three years of student data will minimize measurement error to the point where it is of little concern.[6]

Underestimating the effect of prior achievement will typically advantage schools and teachers with students who have high pretest scores and disadvantage those with students who have lower pretest scores This bias occurs because underestimating the effect of prior achievement produces inflated expected scores for traditionally underperforming students and deflated expected scores for higher performing students. Program designers who wish to base educator incentives on value-added estimates should be aware of the potential for measurement error in the pretest to bias results and consider the applicability of potential remedies. The method Chicago, Amphitheater uses, and in the future Hillsborough will use, is relatively easy to apply once the basic value-added system is in place. Technically inclined readers can refer to Meyer's (1992) article and the text on measurement error models by Fuller (1987) for a description of how to implement this.

C. Student Mobility

Student mobility causes difficulty in attributing achievement gain to schools or teachers because students may have attended more than one school or classroom over the period between the pre- and post-test. Attributing students' gain on a post-test to the school or teacher can be both unfair and inaccurate if the student attended the school or class for a relatively short time. Many value-added models simply drop mobile students from the data set used to calculate value-added or specify that these students are included only if they have attended the school or been assigned to the classroom a minimal number of days (e.g., more than half or two-thirds of the school year). Dallas goes as far as including only those students assigned to a teacher at the beginning of the school year or semester who are still with the same teacher when the test is given and have less than a threshold of days absent. While dropping mobile students may be an acceptable solution when making value-added estimates for research purposes, it can be problematic when using them to determine incentives. Dropping these students eliminates the incentive to work with them to improve achievement.

The general solution used in three of TIF value-added models is to conceptualize the time a student is assigned to a teacher or school as a "dose" (analogous to a dose of medicine) and to include some measure of that dose in the model. For example, the model Mathematica uses allocates 'credit' to a school or teacher based on the fraction of the time during the school year for which the student was assigned. The EVAAS model uses a similar method in which the fraction of time a student is assigned to a teacher or school (e.g., 0.80 instead of 1) represents the teacher or school in the model. Note that when dealing with a single school year, the teacher or school effect is essentially the weighted average of the estimates for each student, with the proportion of the school year as weights. This allows different schools or teachers to get credit for a particular student's gain when multiple teachers or schools are involved. The model used in Chicago also uses a dose model that apportions the growth of students between tests across schools and classrooms in proportion to the time a student spent in the

6 The full EVAAS multivariate model structure does not use pretest scores to predict post-test scores. The structure treats all of the test scores as measured with error, and the error is "absorbed" into the error of the value-added estimate. More years of test scores tend to reduce this error.

school or classroom. This model also controls for the number of times a student changes schools. However, for a large district like Chicago, it is fairly time consuming to assemble the data file needed to estimate a dosage model.

D. Mid-Year Testing

Because most incentive systems use performance over a school year as a base, testing that does not align with the school year causes problems of productivity attribution. Incentive systems often present expectations for schools or teachers as achieving one year of growth. But when testing is done in early March or prior, the learning that value-added measures has actually occurred over a time period when the student likely had two different teachers and may have attended two different schools. The latter is common when students transition between elementary and middle and middle and high school. Talking about value-added as measuring a year's worth of growth loses some of its intuitive appeal, and the SEA or LEA must decide how to attribute the value-added estimate to one school or teacher.

While the logical remedy for mid-year testing would be to test as near to the beginning or end of the year as possible, many states have not taken this step. One method of addressing this problem involves replacing school or teacher indicators in the model with the fraction of the school year each teacher or school was responsible for the student. For example, if the testing is in late February, Teacher A is responsible for instruction from March through June (about 36 percent) and teacher B for September through February (about 64 percent). This method essentially estimates each teacher or school effect as the total value-added estimate multiplied by the appropriate fraction of the school year (i.e., between the pretest and the end of the year in one year and the beginning of the next school year and the post-test). It assumes that teachers or schools contribute to the total value-added in proportion to the time they were responsible for the student. If two successive yearly value-added estimates are available, programs can construct a school year teacher or school effect by calculating a weighted average of the two value-added estimates that collectively cover the school year (the weights being the fraction of the first school year from the test period to the end of that school year and the fraction of the second school year from the beginning of the school year until the next test date).

If there are three years of student data and substantial movement of students from teacher to teacher each year, programs can obtain separate value-added estimates for each teacher or school for each year. This is possible because one can compare the gains for students with different combinations of teachers or schools. For example, in a classroom value-added model, if a substantial number of students from teacher A's class move out to several different classrooms, the effect of Teacher A's instruction should show up as a difference in gain between (a) the students s/he taught who moved to different classes and (b) the students s/he did not teach who ended up in these same classes (i.e. students who were taught by different teachers in the prior year). Models can make similar estimates at the school level when there is substantial mobility between schools. But if almost all fifth graders from elementary school A continue to sixth grade at middle school B, no model can estimate these effects precisely because the model needs to be able to compare the achievement of students who moved from A to B with those who moved from A to middle school C and with students who moved from other elementary schools to either B or C. With only movers between A and B, the effect of A cannot be separated from the effect of B.

E. Value-Added Estimates for Small Schools or Classrooms

Because the precision of value-added estimates depends greatly on the number of students tested from within a school or classroom, estimates for small schools or classrooms tend to be less precise. Researchers sometimes express precision as the width of a confidence band around the estimate that quantifies the range of values in which the estimate might fall. The lower the precision in a model, the wider the range of potential value-added estimates. Thus, if there are a lot of small schools or classrooms, it will be hard to be sure that teachers or schools with different value-added estimates truly differ from one another. Because of the smaller number of students that is the basis for the estimate, schools or teachers with a smaller number of students also tend to show up more often at the top or bottom of the value-added distribution and are also more likely to change their position in the distribution over time (Goldhaber & Hansen, 2008; Kane & Staiger, 2002). These factors make program administrators less confident in rewarding some educators and not others based on value-added scores.

Value-added models TIF 1 and 2 grantees used have usually addressed this problem in one of two ways: through shrinkage or by increasing the number of years of data used in the model.

Shrinkage moves the estimate of a school's or classroom's value-added toward the mean value-added score of other schools or classrooms in the state, region, or district. A "shrunk" estimate is based on a weighted average of: (a) the estimate for the specific school or classroom and (b) the average estimate for all schools or classrooms, with the weights based on the precision of each estimate. Since smaller units have lower precision, all else equal, small schools or classrooms move closer to the overall average. This counteracts the tendency of small units to fall at the top and bottom of the distribution and makes their position more stable from year to year (see McCaffrey, Han, & Lockwood, 2008).

Work by Mathematica Policy Research (Potamites, Booker, et al., 2009, Potamites, Chaplin, et al., 2009) found that shrinkage reduced the error of school value-added estimates, though it also reduced the overall variation between schools, making it harder to distinguish differences between schools. Researchers at the VARC have also found that shrinkage reduces variance at the ends of the value-added distribution and that small schools appear less frequently as outliers. A consensus in favor of shrinkage is beginning to appear in value-added literature. The EVAAS, Chicago, Dallas, and Mathematica models all use some form of shrinkage. Value-added estimates after shrinkage are likely, on average, to be more accurate representations of true productivity. They are also likely to be fairer to educators in small schools or who teach few students.

The second way to address the precision problem is to increase the number of years of data used in the estimate. In essence, this approach increases the sample size by using student data from multiple years. Since the model includes more years of data, the precision of the model improves for all estimates. There are a number of ways to do this, and providers of value-added estimates to TIF grantees have used a few different methods. Mathematica (Potamites, Booker, et al., 2009; Potamites, Chaplin, et al., 2009) used two years of data to estimate value-added for the Memphis and EPIC grantees. The VARC has also worked with multiple years of data. Another approach would be to estimate two or three years of value-added separately and then use a weighted average for rewards.

It appears that adding years of data substantially increases the precision of school and classroom

estimates. The Mathematica researchers reported a 20 to 40 percent decrease in the proportion of variation among schools attributed to measurement error. These researchers also found that adding another year of data had a much greater effect than shrinkage on reducing the standard error of school effect estimates. Goldhaber and Hansen (2008) found that using two or three years of data in classroom value-added estimates also increased precision, thus allowing a greater distinction of between 2 to 7 percentage points from the SEA, LEA, or school average.

There are two potential drawbacks to using multiple years of data. The first is that there may not be enough years of data for all teachers, either because of a lack of student data or because the teacher has not taught for more than one year. The second drawback is that averaging across years obscures true changes. For example, if a teacher improves substantially in his/her second year, he/she may still be ineligible for an incentive if his/her first year value-added was low. Thus, for teachers or schools with low value-added scores, this could make them relatively less motivated by the incentive to improve because they would fail to receive an award despite the improvement in student achievement. In addition, requiring multiple years of data delays the payment of the first incentive until the SEA, LEA, or school has accumulated the appropriate years of data.

Having multiple years of data is probably less an issue when SEAs and LEAs use value-added estimates to award annual bonuses. Mistakenly awarding a one-time bonus to educators whose value-added estimates were in error would not have a lasting effect since the educator would most likely not receive a bonus the following year. Similarly, when using value-added to transfer or terminate educators, multiple years of data would make the action more defensible.

F. Using Value-Added When There Are Only a Few Schools or Classrooms to Compare

Problems may arise when there are only a relatively small number of schools or classrooms among which SEAs and LEAs compare value-added. First, in a small comparison group, it may be hard to tell if a particular level of value-added really represents high or low productivity. While one might find that school A has the highest productivity in a group of six schools, the average of the six might be relatively low when compared to a larger and more representative group of schools. Second, the smaller the comparison group (of schools or teachers) for which value-added is estimated, the more the results for any teacher or school depend on the results from the individual schools or teachers in the comparison group. For example, because value-added estimates are typically relative to the group mean, with fewer schools or classrooms the mean depends more on each school or classroom score. Thus, if one school or classroom does particularly well or poorly, this will have a greater effect and therefore lower or raise the mean value-added score for the entire control group.

One solution to both of these problems is to move to a larger comparison group, which involves linking up with a larger measurement system. For example, Weld County/Fort Lupton decided to wait to use a growth measure until the state finished building its own measurement system that would estimate expected performance for each school in the state. Weld County/Fort Lupton then used the state calculations of school performance, relative to all other schools in the state, as one performance measure for its school award. Similarly, the PICCS charter school consortium in New York City is considering having value-added estimates made using the same system as the New York City school

district. This would allow the PICCS to compare the value-added of each school with similar New York City public schools. Last, some TIF grantees in states using the EVAAS model already have value-added estimates based on a comparison of schools or teachers statewide. In all of these cases, because the comparison group is larger, the influence of any one school in the incentive program is relatively low. Thus, the comparison standard (mean performance in the larger group) is more meaningful, and it is easier to determine the true value-added of the grantee's schools or classrooms. As more states move to developing statewide value-added systems, it will be more attractive for TIF grantees with relatively few schools to link up with the state system to produce a more accurate performance measure.

Another way to address these problems is to define performance expectations using a value-added model for a set of base years. The idea is that programs can develop a value-added model based on prior years' data, and the coefficients from that model (i.e., the pretest coefficient as well as those for any student characteristics in the model) define a benchmark that represents average productivity at that point in time. This model defines the expected relationship between post-test and pretest, or the expected productivity. For subsequent years, grantees can derive a school's or teacher's predicted achievement by multiplying the regression coefficients from the baseline model with the pretest scores and demographics of the current students. Once the SEA, LEA, or school determined the predicted value-added score, it would subtract the predicted score from the actual post-test scores of the current students. The SEA or LEA would consider schools or teachers with positive values doing better than baseline expectations and those with negative values worse. Arizona's Amphitheater school district uses this approach in its PBCS.

Furthermore, when using this approach, the grantee is not comparing schools or teachers with each other, but with past average performance. The efforts of one school or teacher do not affect whether another earns an incentive, and in theory all schools or teachers could have positive value-added (if all had higher productivity than was typical in the base line years). If four years of student test scores are available, one could average across three sets of model coefficients to get predictors that would likely be quite reliable because of the larger sample of students used to calculate the predicted scores. On the other hand, there are two potential limitations to this approach. First, if tests change between the current and baseline years, the average productivity represented by the baseline model will no longer be a good reference point. Meyer and Dokumaci (2009) found that the tests in the state they studied were not always comparable from year to year. If grantees use state standardized tests, they should re-estimate the baseline model every few years. Second, because it is possible for all schools or teachers to exhibit positive value-added, grantees may have a more difficult time projecting incentive costs than if they had based incentives on relative value-added scores. In the latter case, the grantee typically would set a threshold for receiving an incentive at some percentile or standard deviation above the average value-added, which allows a relatively close estimate of how many schools or teachers will receive payments. (For example, if teachers with value-added one standard deviation above the mean qualify, and value-added has an approximately normal distribution, then about 16 percent will receive payments.)

How to Set Cutoff Points for Bonus Payments

Designers of any incentive system based on measures of performance need to set a threshold level of performance that all eligible staff must meet in order to earn the incentive. Program planners need to ask, how much value-added is needed to earn the incentive? Due to the limited experience with value-added, it can be hard to answer this "how much" question. It is harder to find a set of obvious reference points compared to attainment standards like meeting or not meeting NCLB's AYP requirement. Among TIF 1 and 2 grantees, the most common approaches are setting thresholds relative to the average value-added and setting them based on percentiles of the value-added distribution.

Several TIF grantees use the average value-added (typically represented by zero in most models) as the reference point. The rationale here is that if teachers or schools are contributing the average value-added, they are succeeding in helping students make one year of growth. This is reasonable when comparing a large number of schools or teachers because the assumption is that the average growth shown by a large group of teachers or schools represents what students should be able to learn in a school year. The incentive design then rewards any teacher or school with value-added at the average level or above. A big advantage to this approach is ease of projecting costs. By knowing the incentive amount and the number of teachers eligible to receive the award (because about half will be at or above average), it is easy for the SEA, LEA, or school to determine how much its PBCS will cost.

A refinement on this approach is to recognize that individual school or teacher value-added estimates likely contain some error, so that a teacher or school just below the average may really have average, or maybe even somewhat above-average, productivity. Amphitheater Arizona TIF takes this problem into account by adding one standard error[7] to each teacher's or school's value-added estimate. This reduces the chances that a teacher or school with a performance that is truly at least average will miss out on a bonus. Amphitheater's incentive structure then provides for three incentive levels: 20 percent of the maximum bonus for a value-added score 0.5 standard deviations[8] below the baseline up to the baseline, 60 percent for a score between the base line and 0.25 standard deviations above the baseline, and 100 percent for a score 0.25 standard deviations above the base line. This structure spreads out the incentive funds relatively widely so that some level of award is likely to be attainable by most teachers.

Some grantees, primarily using the TAP model, have used a similar approach for individual school or teacher performance awards. This model measures performance levels based on the number of standard errors each school or teacher value-added estimates deviates from the mean. A school or classroom is considered average if its value-added estimate falls within one standard error (plus or minus) of the mean. Generally, this cutoff defines the first level of incentive payment in districts with this type of model. Grantees consider schools or classrooms with a value-added estimate one standard error above the mean to be above average and usually award a larger amount than to those schools deemed to have average performance. Schools or teachers receive the largest incentive amount when the value-added is two standard errors or more above average, a level defined as well above average effectiveness. Based on standard statistical theory, teachers or schools

7 The standard error is a measure of the uncertainty of an estimate like an average. It represents the average amount that an estimate like an average might change in different samples of the same size due to picking different samples from the same population.

8 The standard deviation is a measure of the dispersion or variability of performance. In essence, it is the average difference from the mean in the sample of schools or teachers for which value-added estimates are available.

at that level are highly unlikely to really be average performers misclassified as above average.

In contrast to using average productivity as the reference point, other grantees have started with school or teacher productivity rankings, typically converting the value-added estimates into percentiles, ranking schools or teachers, and providing those at the top of the ranking with the bonus. For example, Houston's ASPIRE program provides teachers or schools in the top quartile of the distribution a full award, and those teachers in the second quartile a partial award. Lower ranked schools or teachers do not receive an award. This essentially allows one-half of teachers to receive an award. Dallas also ranks schools and teachers using value-added, then divides the ranks into three award categories: the 90th to 99th percentile, the 80th to 89th percentile, and the 70th to 79th percentile. Those below the 70th percentile do not receive an award. The Dallas structure provides for a higher degree of differentiation, in that only 30 percent of schools and teachers are likely to receive any bonus, and there is further productivity-based differentiation within that 30 percent. In both these cases, these thresholds were determined based on the amount of funding available, the bonus amounts the district felt were meaningful, and the philosophy that the grantee should reward only better than average performance.[9]

The "top down" approach illustrated by Dallas and Houston has the advantage of easy cost estimation because it keeps costs and bonuses stable. Predetermining the percentile cutoffs allows PBCS designers to determine how many teachers or schools will receive each incentive level (e.g., 30 percent in Dallas). In contrast, grantees cannot determine standard errors until after the data are in, so it is not so easy to estimate costs.[10] On the other hand, the choice of percentiles for cutoffs in the top down approach used by Houston and Dallas might seem arbitrary to educators. The use of "average" value-added as the reference point and standard deviation or standard error-based cutoffs provides an empirically determined basis for the thresholds.

In both of these approaches, the number of incentive levels is an important design decision. Having multiple levels allows for substantial differentiation based on performance, while also providing achievable performance goals for teachers or schools across the performance distribution. For example, teachers or schools with below-average performance can more realistically aim for average performance than for performance at the 75th percentile. Similarly, teachers or schools with average performance can more realistically aim for performance 0.5 standard deviations above the mean than for performance standard deviation above. Since achievable incentives are more likely to motivate effort than unachievable ones, a set of graduated performance thresholds will likely have a positive motivational impact on more educators. Graduated thresholds will allow for small bonuses that many educators can attain, thus improving buy-in while still providing substantial differentiation.

Setting thresholds based on standard errors helps PBCS designers differentiate the award amounts and plan a budget for their PBCS. Assuming that productivity variation is approximately normally distributed, setting thresholds based on standard errors allows SEAs, LEAs, and schools to determine a fairly accurate estimate of the number of schools or teachers likely to earn each bonus level. (For example, in the interval between 0.25 and 0.5 standard deviations above the mean, one would

9 Dallas has also used a four-level structure with percentile thresholds at 60-69, 70-79, 80-89, and 90-99.

10 While a few years of experience may provide a good basis for projecting standard errors, this can be complicated by the standard error getting smaller if the bottom schools improve.

find about 9.3 percent of teachers.) In order to set meaningful thresholds, SEAs, LEAs, and schools would need to set attainable and worthwhile goals that take into consideration the amount of error in teacher or school value-added estimates.

When setting performance thresholds, one could start by considering thresholds such as 0.5, 1, and 1.5 standard deviations from the average value-added. A threshold of 0.5 standard deviations above average corresponds to the difference between the 50th and 69th percentile teacher or school, 1 standard deviation is the difference between the 50th and 84th percentile, and 1.5 standard deviations is the difference between the 50th and 93rd percentile teacher or school. These thresholds would provide significant opportunities to differentiate the award structure. For example, under this type of model, a grantee could provide first-level bonuses to about 15 percent of eligible recipients, second-level bonuses to about 9 percent of participants, and top-level bonuses to about 6 percent of eligible recipients.

Differences between thresholds should represent educationally meaningful differences in student outcomes. That is, the difference between the minimum level needed to earn a bonus and the average value-added should represent a substantial increment in student achievement. One way to check this is to compare the difference between value-added cutoff points for different bonus levels to the average grade-to-grade gain in a subject (which is likely to vary by test and subject.). This threshold is easiest to calculate and understand when tests are vertically scaled (i.e., have a meaningful point scale that spans multiple grades). Schochet and Chiang (2010) report an average gain of 0.65 standard deviations in reading and 0.94 in mathematics across four upper elementary grades for seven standardized tests. This translates into about 3 standard deviations in teacher value-added for reading and 4.4 in math.[11] Meyer and Dokumaci (2009) found that one year of scale score growth averaged about 2.5 standard deviations of school value-added in reading and 3.3 in math for one state's grades 3-8 assessments. Using Schochet and Chiang's numbers for teachers, differences in thresholds of 0.5 a standard deviation above average value-added would represent about one-sixth year of growth in reading, 1 standard deviation about one-third year, and 1.5 standard deviations two-thirds of a year. Using Meyer and Dokumaci's school numbers, 0.5 standard deviation would represent about one-fifth of a year, 1 standard deviation about two-fifths, and 1.5 about three-fifths of a year growth in reading. These seem like educationally significant differences.

Grantees should also consider the amount of error in the value-added measure when setting performance thresholds. More error increases the likelihood that PBCS will reward some truly below average schools or teachers (in terms of performance). In addition, increased error may also fail to reward school or teachers that are truly above average. Both types of misclassification can be costly to program designers. McCaffery, Han, and Lockwood (2008) observed that repeatedly failing to reward a teacher who is truly high performing could lower that teacher's motivation and lead to turnover, while repeatedly rewarding below-average performing teachers would send the message that they were good performers and do not need to improve. In addition, one might contend that rewarding those teachers who are truly average or below also spends money that the state, district, or school could use elsewhere. McCaffery, Han, and Lockwood (2008) pointed out that when the performance measure has substantial error, a higher threshold for receiving the bonus makes it less likely that the grantee will reward teachers or schools that are not truly above

11 See Schochet & Chiang (2010) pages 21 and 22, and note 4.

average. Thus, if it is important to maximize the degree to which bonuses go to those who are truly above average, then the minimum threshold needs to be set above average. How far above average the PBCS sets the threshold depends on the amount of error, the minimum size of the bonus, and the value placed on avoiding paying bonuses to below-average performers versus mistakenly not paying higher than average performers.

The simulation study by Schochet and Chiang (2010) provides some insight into how measurement error in value-added estimates and choice of threshold levels interact to produce different proportions of teachers misidentified as deserving or not deserving of a bonus. This study found that if one year of value-added estimates were used, given a set of assumptions about measurement error, a threshold equivalent to about one-half of a teacher's value-added standard deviation[12] would result in about 27 percent of the teachers receiving the incentive actually having average or lower true productivity and about 14 percent of those with productivity at or above the threshold missing the bonus. These percentages drop to 25 percent and 6 percent at a threshold equivalent to about .93 teacher value-added standard deviations and 23 percent and 2 percent at a threshold equivalent to 1.4 standard deviations above average teacher value-added. An interesting point here is that moving the threshold up has only a limited effect on the proportion of teachers who receive a bonus but have average or less true productivity. This is largely due to the substantial measurement error in one year of teacher value-added results assumed in the simulation, an assumption based on a review of 25 value-added studies. Thus, if program designers want to ensure that they only reward teachers or schools with above average value-added, the minimum threshold needs to be set quite high. But a minimum threshold as high as 1 standard deviation above average will provide rewards to only about 16 percent of the teachers or schools, assuming value-added estimates are normally distributed.

The alternative to setting a high threshold is using more years of teacher or school data. In their study, Schochet and Chiang showed that moving from one to three years of data is likely to reduce erroneous classification more than raising the threshold from 0.5 to 1 teacher standard deviation. Adding two more years of data resulted in 17 percent of the teachers receiving the incentive having average or lower true productivity, and missing only 8 percent of those with productivity at or above a 0.5 teacher standard deviation threshold. The disadvantages of using more years of data, as discussed in the section on small schools and classrooms, are: (1) multiple years of data may not be available for all teachers, (2) an estimate based on multiple years ignores what may be true changes in year-to-year productivity, and (3) the incentive cannot be paid until enough years of data have accumulated.

Again, this suggests that programs can benefit from using multiple incentive levels. A minimum threshold (such as 0.5 standard deviations above average) can be set at a level that would be attainable by a substantial number of schools or teachers. But the bonus associated with that threshold would be set at the minimum meaningful amount. Though a substantial proportion of the educators receiving the bonus would not be above average performers, the lower bonus amount makes these errors less expensive as well as having positive motivational impact on those with below-average performance by providing an attainable goal. The grantee would provide higher bonus amounts for higher thresholds, which will have a lower probability that below-average teachers or schools could get the

12 The simulation set thresholds in terms of standard deviations of student gain scores rather than of teacher value-added. The researchers converted the thresholds provided into teacher value-added standard deviations using the conversion factor provided on page 22, note 4.

bonus. This method would recognize those likely to have higher productivity and provide a goal for higher performers.

Interestingly, no TIF grantee is using a formula that is a continuous function of value-added, which provides a set amount of dollars per each unit of value-added, thereby making it more difficult to calculate and project an estimated cost, but eliminating the need to set thresholds. South Dakota's Incentives Plus TIF program is considering a concept that is similar to this model. South Dakota's value-added model will project each student's expected score (in norm curve equivalents), and count the number of students whose actual score is above the prediction. The incentive will be pro-rated based on the percentage of a teacher's students who have higher than expected scores. Again, this avoids having to set specific value-added thresholds, but it does set an implicit standard that the program will reward average performance (one year of average growth). If a state, district, or school eventually implements this structure, it will be interesting to see how the amounts and distribution of the incentives compare with other approaches.

Summary

Though the issues involved in using value-added are complex and there is still disagreement among experts about the best way to estimate and use value-added, there is evidence of an emerging consensus. In addition, there are now more alternatives that PBCS designers can use to surmount some common limitations of value-added models. It seems increasingly clear that student demographic controls should be used in models that use only one year of data (i.e., are based on two rounds of testing). It also seems advisable to consider pretest measurement error by either using several years of student data or one of the pretest measurement correction methods discussed above. PBCS designers should consider the "dose" approach for dealing with student mobility or team teaching. When small schools or classrooms are an issue, applying a form of shrinkage to value-added estimates or using multiple years of school or teacher data will likely provide more valid estimates. Where the compensation system will include a relatively small number of teachers or schools, designers should consider broadening the universe of comparison by moving to a statewide or region-wide value-added system. This will become easier as many more states are moving toward developing statewide systems in response to Race to the Top and other Federal incentives. If no such system is available, states, districts, or schools should consider comparing value-added to a historical baseline. Last, designers should consider designing the incentive structure with multiple value-added thresholds linked to graduated incentive amounts.

While most educators are not going to become value-added modelers, value-added is likely to be prominent on the educational landscape for the foreseeable future. Current value-added methods are not perfect measures of productivity, but they are arguably more valid and fair overall than the alternatives. Indeed, value-added estimates of school and teacher performance have become an important prerequisite for Federal, and increasingly state, policy on assessing educator quality, monitoring the distribution of effective educators, and rewarding educator performance. Therefore, learning more about value-added should be a priority for designers of PBCS for educators.

References

Ballou, D. (2005). Value-added Assessments: Lessons from Tennessee. In R. Lissitz (Ed.), *Value-added Models in Education: Theory and Application* (pp. 272-297). Maple Grove, MN: JAM Press.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for Student Background in Value-added Assessment of Teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37–66.

Bembry, K., Weerasinghe, D., & Mendro, R.L. (1997). *Classroom Effectiveness Indices, Statistical Methodology, Post-Hoc Analysis, and Practical Applications*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Illinois.

Braun, H., Chudowsky, N., & Koenig, J. (2010). *Getting Value Out of Value-added: Report of a Workshop*. Washington, DC: National Research Council. Available at: http://www.nap.edu/catalog/12820.html

Fuller, W.A. (1987). *Measurement Error Models*. NY: Wiley.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). *Evaluating Teachers: The Important Role of Value-added*. Washington, DC: Brown Center on Education Policy at Brookings.

Goldhaber, D., & Hansen, M. (2008). *Assessing the Potential of Using Value-added Estimates of Teacher Job Performance in Making Tenure Decisions*. Policy Brief. National Center for Analysis of Longitudinal Data in Education Research. Available at: http://www.urban.org/publications/1001265.html

Hanushek, E.A., & Rivkin, S.G. (2010, January). *Generalizations About Using Value-added Measures of Teacher Quality*. Paper presented at the annual meeting of the American Economic Association, Atlanta, Georgia.

Harris, D. (2011). *Value-added Measures in Education: What Every Educator Needs to Know*. Cambridge: Harvard University Press

Isenberg, E., & Hock, H. (2011). *Design of Value-added Models for IMPACT and TEAM in DC Public Schools, 2010–2011 School Year. Final Report*. Princeton, NJ: Mathematica Policy Research.

Kane, T.J., & Staiger, D.O. (2002) The Promise and Pitfalls of Using Imprecise School Accountability Measures. *The Journal of Economic Perspectives, 16*(4), 91–114.

Kane, T. J., & Staiger, D. O. (2008). *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation*. NBER Working Paper 14607. Cambridge, MA: National Bureau of Economic Research.

McCaffrey, D.F., Han, B., & Lockwood, J.R. (2008). *From Data to Bonuses: A Case Study of the Issues Related to Awarding Teachers Pay Based on Their Students' Progress*. National Center on Performance Incentives Working Paper 2008-14. Nashville, TN: National Center on Performance Incentives.

Meyer, R.H. (1992). *Applied Versus Traditional Mathematics: New Econometric Models of the Contribution of High School Courses to Mathematics Proficiency*. Discussion Paper No. 966-92. Madison, WI: Institute for Research on Poverty, University of Wisconsin-Madison.

Meyer, R.H., & Christian, M.S. (2008, February). *Value-added and Other Methods for Measuring School Performance: An Analysis of Performance Measurement Strategies in Teacher Incentive Fund Proposals*. Paper presented at National Center for Performance Incentives Conference, Performance Incentives: Their Growing Impact on American K-12 Education, Nashville, TN.

Meyer, R.H., & Dokumaci, E. (2009). *Mean and Variance Value-Added Indicators With Multilevel Shrinkage: Application to a Multi-District Statewide Value-Added System*. Paper presented at the American Education Finance Association Annual Conference, Nashville, Tennessee.

Meyer, R.H., & Dokumaci, E. (2010). *Value-added Models and the Next Generation of Assessments*. Paper presented at the Exploratory Seminar: Measurement Challenges within the Race to the Top Agenda sponsored by the Center for K-12 Assessment & Performance Management, Educational Testing Service. Available at: http://www.k12center.org/publications.html

Milanowski, A.T., Witham, P., Schuermann, P., Kimball, S., & Pietryka, D. (2010, March). *Harvesting Lessons on Educator Incentive Plan Design from Technical Assistance Provided to Teacher Incentive Fund Grants*. Paper presented at 2010 Annual Meeting of the American Educational Finance Association, Richmond, VA.

Potamites, L., Booker, K., Chaplin, D., & Isenberg, E. (2009). *Measuring Teacher and School Effectiveness in the EPIC Charter School Consortium – Year 2 Final Report*. Washington, DC: Mathematica Policy Research.

Potamites, L., Chaplin, D., & Isenberg, E., & Booker, K. (2009). *Measuring School Effectiveness in Memphis-Year 2 Final Report.* Washington, DC: Mathematica Policy Research.

Sanders, W.L. (2006). *Comparisons Among Various Educational Assessment Value-Added Models*. SAS Institute. Paper presented at The Power of Two – National Value-added Conference, Columbus, OH. Available at: http://www.sas.com/resources/asset/vaconferencepaper.pdf

Sanders, W.L., Wright, S.P., Rivers, J.C., & Leandro, J.G. (2009). *A Response to Criticisms of SAS EVAAS*. SAS Institute White Paper. SAS Institute. Available at: http://www.sas.com/resources/asset/Response_to_Criticisms_of_SAS_EVAAS_11-13-09.pdf

Schochet, P.Z., & Chiang, H.S. (2010). *Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains* (NCEE 2010-4004). Washington, DC: National Center for Evaluation and Regional Assistance, Institute of Educational Sciences, U.S. Department of Education. Available at: http://ncee.ed.gov

Watson, J., Witham, P., & St. Louis, T. (2010). *Evaluating Student-Teacher Linkage Data in Teacher Incentive Fund (TIF) Sites: Acquisition, Verification, and System Development.* Madison, WI: Center for Educator Compensation Reform & Value-added Research Center.

Webster, W., Mendro, R., Orsak, T., & Weerasinghe, D. (1998). *An Application of Hierarchical Linear Modeling to the Estimation of School and Teacher Effect.* Retrieved from http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED424300&ERICExtSearch_SearchType_0=no&accno=ED424300 on August 6, 2010.

Table A1 – Use of Value-added by Round 1 & 2 TIF Grantees[13]

Grantee	Planned to Use School VA?	Using School VA 09-10?	Planned to Use Classroom VA?	Using Classroom VA 09-10?	Internally or Externally Developed	Comment
Amphitheater	No	Yes	No	Yes	Internal	
Beggs	No	No	No	No		VA judged incompatible w/ state model
CEI-PEA/PICCS	Yes	No	Yes	No		Working toward using VA
CTAC/ Charlotte-Mecklenburg	No	No	No	Yes	Internal	Added for 09-10 school year
Chicago	Yes	Yes	Yes	No	External –VARC	
Chugach	No	No	No	No		Use "value table"
Cumberland	No	No	No	No		
Dallas	Yes	Yes	Yes	Yes	Internal	Have been using some form of VA for 15 years
Denver	No	No	No	No		Colorado Growth model
Eagle County	Yes	No	Yes	No		
Edward W. Brooke	Not Clear	No	Yes	No		
Florence	Yes	Yes	No	Yes	External – SAS	TAP
Guilford	No	No	Yes	Yes	External – SAS	
Harrison	No	No	No	No		Proficiency change model
Hillsborough	No	No	No	No		Planning to move to VA
Houston	Yes	Yes	Yes	Yes	External – SAS	
Lynwood	No	No	No	No		
MIT Academy	No	No	No	No		Inactive
Miami-Dade	No	No	NA	NA		Principals only
New Leaders (DC)	Yes	Yes	No	Yes	External – MPR	In new DC teacher evaluation
Grantee	Planned to Use School VA?	Using School VA 09-10?	Planned to Use Classroom VA?	Using Classroom VA 09-10?	Internally or Externally Developed	Comment
New Leaders (Memphis)	Yes	Yes	Yes	No	External – MPR	
New Leaders (Charters)	Yes	Yes	Yes	No	External – MPR	
NIET Algiers	Yes	Yes	Yes	Yes	External – SAS	TAP
Northern New Mexico	No	Yes	No	No		
Ohio	Yes	Yes	Yes	Yes	External – SAS	TAP
Orange County	No	No	No	No		

13 Information based on grantee self-evaluations, project proposals, and Meyer & Christian, 2008.

Grantee	Planned to Use School VA?	Using School VA 09-10?	Planned to Use Classroom VA?	Using Classroom VA 09-10?	Internally or Externally Developed	Comment
Philadelphia	Yes	Yes	Yes	No	External – SAS	TAP
Pittsburgh	Yes	Yes	NA	Will eventually	External – RAND	Principals only
Prince George's County	No	No	Yes	Will eventually		
School of Excellence in Education	No	No	No	No		Post-pre gain for teachers
South Carolina	Yes	Yes	Yes	Yes	External – SAS	TAP
South Dakota	No	No?	No	Planning for 09-10	Internal	
U of Texas System	Yes	Yes	Yes	No	External – SAS	TAP
Weld/Ft. Lupton	No	No	No	No		Uses Colorado Growth model

The work described in this paper was supported by the U.S. Department of Education through the Center for Educator Compensation Reform. The opinions expressed are those of the authors and do not necessarily reflect the view of the U.S. Department of Education, the Center for Educator Compensation Reform, or the institutional partners of the Center. This is a working paper describing initial results of an ongoing project. Comments and suggestions are welcome.

The Center for Educator Compensation Reform (CECR) was awarded to Westat—in partnership with Learning Point Associates, Synergy Enterprises Inc., Vanderbilt University, and the University of Wisconsin—by the U.S. Department of Education in October 2006.

The primary purpose of CECR is to support Teacher Incentive Fund (TIF) grantees in their implementation efforts through provision of sustained technical assistance and development and dissemination of timely resources. CECR also is charged with raising national awareness of alternative and effective strategies for educator compensation through a newsletter, a web-based clearinghouse, and other outreach activities.

This work was originally produced in whole or in part by the CECR with funds from the U.S. Department of Education under contract number ED-06-CO-0110. The content does not necessarily reflect the position or policy of CECR or the Department of Education, nor does mention or visual representation of trade names, commercial products, or organizations imply endorsement by CECR or the federal government.

Center for
Educator Compensation
Reform

Allison Henderson, Director
Phone: 888-202-1513
E-mail: cecr@westat.com

33407.0811.83670507