In the following report, Hanover Research examines the use of student test scores in teacher evaluations within pay for performance systems. The report includes an overview of the evaluation methods used in these systems, a discussion of system costs, expert critiques of the use and weighting of test scores in teacher evaluations, and a review of the outcomes of pay for performance systems.
# Table of Contents

Executive Summary and Key Findings ................................................................. 3  
   Key Findings .................................................................................................. 3  

Section I: Pay for Performance Systems ......................................................... 5  
   Methods of Evaluation in Selected Pay for Performance Systems .............. 5  
      TAP ............................................................................................................. 5  
      Harrison School District #2 ....................................................................... 7  
      ProComp .................................................................................................... 8  
   Costs of Evaluation Systems ........................................................................ 9  
      TAP ............................................................................................................. 9  
      Harrison School District #2 ....................................................................... 9  
      ProComp .................................................................................................... 10  
      Q Comp ..................................................................................................... 10  

Weighting of Student Test Scores in Teacher Evaluations .............................. 11  

Section II: Evaluating Teachers of Non-Tested Subject Areas ....................... 13  
   School/District-Wide Value-Added Model .................................................... 13  
   Pro-Rated School/District-Wide Value-Added Model ................................... 13  
   New Pre- and Post-Tests for Non-Tested Subjects ....................................... 14  
   Curriculum-Based End-of-Course Assessments .......................................... 15  
   Interim Assessments ..................................................................................... 15  
   Locally-Created Assessments ...................................................................... 16  
   Student Learning Objectives ........................................................................ 17  

Section III: Evaluator Training and Reliability of Evaluation Ratings .......... 21  
   Reliability Versus Agreement ....................................................................... 21  
      TAP ............................................................................................................. 21  
      Q Comp ..................................................................................................... 22  

Section IV: Student and Teacher Outcomes .................................................. 23  
   Effect on Student Achievement .................................................................... 23  
   Effects on Teacher and School Culture ....................................................... 24  

Conclusion ...................................................................................................... 26
EXECUTIVE SUMMARY AND KEY FINDINGS

As a result of the No Child Left Behind Act (NCLB) and subsequent federal and state-wide educational accountability initiatives, policymakers have demonstrated increased interest in sophisticated measures of student achievement. The growing emphasis on these measures, as well as the state and federal funding awarded or denied on the basis of student achievement data, has brought with it efforts to align teacher evaluations and compensation with such data. The vast majority of U.S. states are now engaged in compensation reform initiatives, which often involve incentives tied to student test scores.¹

In order to provide a guide for schools and districts faced with or contemplating such reforms, Hanover Research examines four aspects of performance-based compensation. Each of four the sections of this report covers one aspect of this topic according to the following division:

- **Section I: Pay for Performance Systems** — This section describes several pay for performance systems with a particular emphasis on their methods of evaluation. We discuss the components of evaluation in each system as well as the weight accorded to each component. This section also reviews the costs associated with various evaluation programs, as well as the literature surrounding the weighting of student test scores in teacher evaluations.

- **Section II: Evaluating Teachers of Non-Tested Subject Areas** — In this section we offer a more general review of evaluation models in non-tested subject areas and grade levels. This section also includes a summary chart of features, strengths, and weakness of the models reviewed.

- **Section III: Evaluator Training and Reliability of Evaluation Ratings** — Section III examines the training required in two pay for performance systems and the protocols that each system has established to ensure consistency among evaluators.

- **Section IV: Student and Teacher Outcomes** — In this section, we discuss the outcomes associated with pay for performance systems. Student achievement outcomes and teacher attitudes receive special attention.

KEY FINDINGS

- All but one of the pay for performance systems reviewed in this report include student test scores as a component of teacher evaluations. In these cases, the value-added data derived from these scores, rather than the scores themselves, enter into the calculation of the teacher’s rating, usually with a weighting of 50 percent. The one exception to this norm, Denver’s ProComp, provides distinct monetary incentives for observation-based evaluation ratings and student value-added measures.

No consensus exists among experts about the appropriate weight to accord student achievement data in a teacher evaluation. Governmental and policy organizations caution against relying heavily on test scores to make critical personnel and compensation decisions. In contrast, the Gates Foundation has published research showing qualified support for some weighting systems, including systems that accord weights of more than 50 percent to student achievement data.

Recent pay for performance initiatives have relied on two primary sources of funding. The first source, state and federal grants, such as Teacher Incentive Fund grants from the U.S. Department of Education, generally has funded experimental ventures such as the TAP system. The second primary source, taxpayer funding, has supported longer-term endeavors such as Denver’s ProComp and Minnesota’s Q Comp.

There appear to be significant costs associated with the robust data systems necessary to provide value-added measurements. Materials for the TAP system report that these costs may include additional testing expenses, annual expenses for the calculations of value-added scores for both district schools and control schools outside the district, and the purchase of a new data management system.

Both ProComp and Q Comp require regular training and annual re-certification for evaluators who conduct classroom observation. The objective of these training sessions is to ensure inter-rater agreement among the evaluators at each school.

Several recent studies of pay for performance initiatives have found that performance pay has a negligible impact on student achievement. Notably, the one experimental study conducted thus far, an investigation of Nashville’s POINT performance-based compensation system, found no impact on student achievement.

Pay for performance systems have had varied effects on teacher attitudes. Despite recurring findings that teachers found the pay for performance systems to be fair, most studies found that teacher practices remained unaltered in response to pay for performance.
SECTION I: PAY FOR PERFORMANCE SYSTEMS

This section describes several pay for performance systems, giving special emphasis to the methods of evaluation used in each system. We discuss the components of evaluation in each system as well as the weight accorded to each component. This section also reviews the costs associated with various evaluation programs, and discusses the literature surrounding the weighting of student test scores in teacher evaluations.

Before proceeding further, it is important to note that the primary achievement data used by many pay for performance systems is value-added data. These data are obtained through value-added modeling, or a process that calculates the amount of growth that a student experiences during a year. This growth usually is based on pre- and post-test scores. While value-added modeling has the advantage of recognizing the growth that teachers foster rather student attainment levels, which may reflect preexisting differences, there are nonetheless some concerns with the use of value-added data as the sole measure of student achievement. This topic shall be reviewed in greater detail later in this section.

In general, when discussing “student achievement data” or “student test scores” throughout the remainder of this report, we will be referring to value-added data derived from standardized tests.

METHODS OF EVALUATION IN SELECTED PAY FOR PERFORMANCE SYSTEMS

TAP

TAP: The System for Teacher and Student Advancement, is a framework for teacher evaluation and compensation developed by the National Institute for Excellence in Teaching (NIET). At present, approximately 500 schools in 17 states employ the TAP system.²

The TAP evaluative process includes two primary components. The first component is the teacher’s demonstration of the appropriate Skills, Knowledge, and Responsibilities (SKRs), which are determined using the TAP Rubrics and Responsibility Survey. This survey is completed four to six times per year for each instructor by TAP-certified administrators, master teachers, and mentor teachers, and includes 19 indicators designed to evaluate a teacher’s overall performance. For this component of the TAP evaluation, instructors receive a score from one to five in each of the areas listed in Figure 1.1 on the following page.³

### Figure 1.1: Skills, Knowledge, and Responsibilities Categories Employed by TAP

<table>
<thead>
<tr>
<th>Instruction</th>
<th>The Learning Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>➤ Standards and Objectives*</td>
<td>➤ Expectations*</td>
</tr>
<tr>
<td>➤ Motivating Students*</td>
<td>➤ Managing Student Behavior*</td>
</tr>
<tr>
<td>➤ Presenting Instructional Content*</td>
<td>➤ Environment*</td>
</tr>
<tr>
<td>➤ Lesson Structure and Pacing*</td>
<td>➤ Respectful Culture*</td>
</tr>
<tr>
<td>➤ Activities and Materials*</td>
<td></td>
</tr>
<tr>
<td>➤ Questioning*</td>
<td></td>
</tr>
<tr>
<td>➤ Academic Feedback*</td>
<td></td>
</tr>
<tr>
<td>➤ Grouping Students*</td>
<td></td>
</tr>
<tr>
<td>➤ Teacher Content Knowledge*</td>
<td></td>
</tr>
<tr>
<td>➤ Teacher Knowledge of Students*</td>
<td></td>
</tr>
<tr>
<td>➤ Thinking*</td>
<td></td>
</tr>
<tr>
<td>➤ Problem Solving*</td>
<td></td>
</tr>
<tr>
<td><strong>Designing and Planning Instruction</strong></td>
<td><strong>Responsibilities</strong></td>
</tr>
<tr>
<td>➤ Instructional Plans</td>
<td>➤ Staff Development**</td>
</tr>
<tr>
<td>➤ Student Work</td>
<td>➤ Instructional Supervision**</td>
</tr>
<tr>
<td>➤ Assessment</td>
<td>➤ Mentoring**</td>
</tr>
<tr>
<td><strong>Expectations</strong></td>
<td><strong>Community Involvement</strong></td>
</tr>
<tr>
<td><strong>Managing Student Behavior</strong></td>
<td><strong>School Responsibilities</strong></td>
</tr>
<tr>
<td><strong>Environment</strong></td>
<td><strong>Growing and Developing</strong></td>
</tr>
<tr>
<td><strong>Respectful Culture</strong></td>
<td><strong>Reflecting on Teaching</strong></td>
</tr>
</tbody>
</table>

Source: University of Indianapolis

*Denotes criteria that are evaluated during classroom observations.

**Denotes criteria that apply only to master and mentor teachers.

The remaining **50 percent of the evaluation score reflects value-added achievement data.** The precise calculation of this portion of the score differs according to whether there are individual student achievement data for a teacher’s classes. When such data exist for a given teacher, 30 percent of the evaluation score depends on individual student achievement data and 20 percent of the score depends on school-wide achievement data. When there are no individual achievement data for a teacher, 50 percent of the total score depends solely on school-wide achievement data. Figure 1.2 depicts this weighting of evaluation scores within the TAP system.

---


The assessment used to provide student achievement data may vary by state or district. TAP schools in Indiana, for instance, obtain their student achievement data from the Indiana Growth Model, a model of student improvement on the state’s ISTEP+ exam. In South Carolina, TAP schools appear to have a variety of options for testing, including the Palmetto Assessment of State Standards and the Measured Academic Progress exam.

**HARRISON SCHOOL DISTRICT #2**

In 2010-2011 Harrison School District 2 adopted a new compensation plan designed by district superintendent Mike Miles, founder of the Focal Point educational leadership organization. In keeping with newly passed legislation in Colorado, the district’s plan bases 50 percent of a teacher’s evaluation score on student achievement data. The district uses a wide-range of achievement data for all subjects, including non-tested subjects. Each teacher has an achievement template crafted specifically for his or her subject and grade level that determines the both the weighting and scoring criteria for each component of achievement. This template may include such components as state tests, district semester exams, district quarterly exams, school-wide state test results, and other individually set goals. Notably, state test results will never account for more than one eighth of a teacher’s overall evaluation score on the basis of this template. Furthermore, the wide range of template designs (86 total for all grade levels and subjects) ensures that teachers in non-tested subjects are measured by standards appropriate to their disciplines. An example of the weighting employed in the template for fourth and fifth grade teachers is presented in Figure 1.3.

---

6 Ibid.
The remaining 50 percent of the score is determined by performance appraisals conducted by the principal. The appraisals for non-tenured teachers include 16 15-minute “mini-observations” and two class-length observations; tenured teachers undergo eight mini-observations and one full-length observation. On the basis of the observations, all teachers receive scores ranging from one to seven in seven performance categories.10

**PROCOMP**

ProComp, a pay for performance system instituted in Denver Public Schools (DPS) in 2006, stands out as one of the first compensation models to abandon the traditional salary schedule altogether. Although the system provides rewards for attaining certain student growth objectives measured by test scores, test data are contractually barred from factoring into the ProComp evaluation. Rather, evaluations are based on observations, interviews, records of teaching or service, and written communications.11 Teachers may earn base-building salary increases for successful ratings on their evaluations and a variety of base-building and non-base-building incentives on the basis of various measures of student growth. Instructors in non-tested subjects are not eligible for (non-base-building) bonuses linked to state test results, but they are eligible for awards based on other student growth objectives and school-wide growth.12

---


COSTS OF EVALUATION SYSTEMS

TAP

The TAP website states that system costs from $250 to $400 per pupil per year to implement and sustain. The Department of Education for the State of Louisiana, which oversees 68 TAP schools, estimates that the costs of the program are $300 to $400 per pupil annually, depending on the structure of the school. The state includes in these costs base salary increments, performance awards, professional development expenses, hiring specialists, and purchasing and maintaining data management and analysis tools.13

The TAP Implementation Manual provides a more detailed breakdown of costs associated with testing. Additional testing, if needed for calculations of student achievement, will represent one such cost. Furthermore, districts will need comparative analyses of value-added data for district schools and control schools. The manual estimates that districts will pay $2.00 per student for value-added calculations for TAP schools and another $2.00 per student for value-added calculations at control schools. Finally, the data management system required for TAP will cost approximately $2,000. These costs are displayed in Figure 1.4. Other TAP expenditures include mandatory TAP training for the leadership team, startup workshops, and travel to sites that have implemented TAP.14

Table 1.4: Additional Costs Associated with Testing in the TAP System

<table>
<thead>
<tr>
<th>Category</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additional Testing</td>
<td>$\text{(Total annual cost of additional testing)} \notag</td>
</tr>
<tr>
<td>$\text{(# of students added to testing program}$\notag x $\text{(cost of test per student)} \notag = \text{(Total annual cost of additional testing)}</td>
<td></td>
</tr>
<tr>
<td>Value-Added Calculations for TAP Schools</td>
<td>$\text{(Total annual cost of value-added calculations)} \notag</td>
</tr>
<tr>
<td>$\text{(# of students x $2.00 per student)} \notag = \text{(Total annual cost of value-added calculations)}</td>
<td></td>
</tr>
<tr>
<td>Value-Added Calculations for Control Schools</td>
<td>$\text{(Total annual cost of value-added calculations)} \notag</td>
</tr>
<tr>
<td>$\text{(# of students x $2.00 per student)} \notag = \text{(Total annual cost of value-added calculations)}</td>
<td></td>
</tr>
<tr>
<td>Data Management System</td>
<td>$\text{Approximately $2,000} \notag</td>
</tr>
</tbody>
</table>

Source: NIET15

HARRISON SCHOOL DISTRICT #2

The Harrison School District implemented its pay for performance system with the aid of an $800,000 grant from the Daniels Fund. Startup costs included both additional salary for 80 percent of the school staff, who merited the increases through their effectiveness ratings, and compensation for new in-house assessment professionals and outside consultants. Additional salary awarded for effectiveness ratings totaled $300,000, and the cost of new hires and consultants for assessment purposes amounted to $200,000. In addition, the costs of developing testing and scoring systems totaled $400,000.16

15 Table text taken verbatim from ibid.
The district claims that its plan is sustainable as long as it maintains an annual promotion rate of 20 to 25 percent. It aims to achieve this goal by making promotions easily achievable for beginning instructors and much harder for more experienced teachers. Moreover, the compensation scale does not include automatic cost of living adjustments, though base figures will receive a triennial review. Because the description of the system does not mention any additional taxpayer or philanthropic funding as a prerequisite for sustainability, it appears that the system will be sustained entirely through the reallocation of existing funds. However, the district also recently applied for and won a $9 million, 5-year TIF grant for professional development initiatives. From the award announcement it is not clear whether this grant was necessary to sustain the program.

**PROCOMP**

The ProComp system may present the most clearly sustainable model of performance-based compensation. After receiving several million dollars in philanthropic funding for research, development, and implementation costs, DPS now funds ProComp through a newly instituted annual property tax that provides $25 million in inflation-adjusted revenue. In addition, after implementation DPS continued to develop its ProComp Assessment Profile, a tool to assess student growth, with the aid of a $500,000 Daniels Fund grant.

**Q COMP**

Q Comp, a pay for performance system adopted by many Minnesota school districts, relies on annual funding of $260 per pupil. Of this $260, $190 comes from state aid, and the remaining $70 is derived from local levies. The recent implementation of Q Comp in Anoka-Hennepin, Minnesota’s largest school district, is expected to cost $10 million annually: $7 million in state aid and $3 million in district tax levies. In order to accommodate the new levy within its current taxation scheme, the district plans to reduce taxpayer funding for other programs.

---

17 Ibid., 19.
20 Ibid., p.6.
WEIGHTING OF STUDENT TEST SCORES IN TEACHER EVALUATIONS

There does not appear to be any research demonstrating a connection between using a specified percentage of student standardized test scores in teacher evaluations and improved student achievement on standardized test scores. To the contrary, while many states have implemented evaluation systems that include data from student test scores, these efforts do not appear to be research-based. A 2011 report by the National Research Council, for instance, concluded that “standardized test scores are of limited value in determining causes of improvements in student performance.” The NRC further states that

The standardized test scores that have been trumpeted to show improvement in the schools provide limited information about the causes of improvements or variability in student performance. This would be true, presumably, for any school system that use standardized tests as a measure of achievement.

If standardized test scores are only of limited value in determining causes of improvements in student performance, then they are even less valuable a factor when assessing teacher performance. Nevertheless, interest in linking a specified percentage of student standardized test scores to teacher assessments, which underlies performance-based compensation, is gaining traction throughout the United States. In 2009, Michelle Rhee (then Chancellor of Washington, D.C. Public Schools) introduced IMPACT—a teacher evaluation system in which 50 percent of D.C. public school teacher evaluations were based on standardized student test scores. Recently, in August 2012, Chancellor Kaya Henderson announced that this percentage would be reduced to 35 percent. Neither Rhee nor Henderson has cited research that validates the effectiveness of either percentage.

Several policy organizations have recommended against using inflexible weighting systems in personnel decisions. For example, in a 2011 review of teacher evaluation systems the Brookings Institute states that they “do not recommend that states or the federal government be prescriptive about the components that districts should include in their teacher evaluation systems or how they should be weighted or how the information from those systems should be used for high-stakes personnel decisions.”

Similarly, the Center for Educator Compensation Reform recommends that

[Value-added measures of student achievement should not be the sole measures that states and districts use to evaluate teacher productivity. Multiple measures of student achievement and multiple assessments of teacher performance are recommended.]

---

24 Strauss, V. “Rhee’s teacher evaluation system is revised—but is it improved?” Op. cit.
Thus, a considerable number of experts are hesitant, at the very least, about assigning specific and significant evaluative weights to test score data.

A somewhat different perspective emerges from the publications of the Gates Foundation. Although the foundation researchers do not endorse any specific weighting scheme explicitly, they list the following as one of the implications of their research: “Whenever multiple measures are being used, policymakers must choose how much weight to attach to each. There are trade-offs, in terms of both validity (in predicting other outcomes) and reliability.”27 In this statement the researchers are referring to their finding that that equal weighting for observations, value-added, and student feedback (a 33 percent weighting for each component) maximizes reliability but lowers predictive power for a teacher’s underlying value-added. Unequal weightings, in contrast, maximize predictive power for underlying value-added but decrease reliability. Importantly, the researchers note that the optimal weighting for predictive purposes, though always dominated by student achievement data, varies depending upon the observation system used for the rating. Thus, as the observation systems changes, the weighting of value-added data that maximizes predictive power ranges from 58 percent to 76 percent.28

Although the foundation’s research could be used to argue for test score weightings of greater than 50 percent, it is important to bear in mind three caveats:

- The Gates Foundation researchers remain unsure that their methodology has definitively identified the correct measures of underlying value-added
- Underlying value-added is simply the ability to improve performance on certain tests; it is not the sole measure of effective teaching
- The Gates Foundation researchers discovered that weightings in which test scores dominated were not more highly correlated with student effort, student emotional attachment, and performance on assessments other than state tests than equal weightings.29

Accordingly, the foundation’s report does not provide unequivocal support for evaluations that place heavy emphasis on test scores.

---

28 Ibid., pp. 50-51, 64
29 Ibid., pp. 52-54.
SECTION II: EVALUATING TEACHERS OF NON-TESTED SUBJECT AREAS

The evaluations used in the systems described in Section I represent only a few of the means of evaluating non-tested subjects. In this section we offer a more general review of evaluation models in these disciplines. Figure 2.1 on page 19 provides a summary chart of features, strengths, and weakness of these models.

SCHOOL/DISTRICT-WIDE VALUE-ADDED MODEL

Many districts and states have adopted the approach of basing the achievement portion of a teacher’s evaluation on school and/or district-wide performance on state standardized exams, particularly for those teachers who teach otherwise non-tested subjects such as art, physical education, and music. The rationale used by many states and districts that have adopted this method is that it builds a “collective school-wide or team-based effort around student achievement.” In other words, districts and schools believe that all teachers should make reasonable contributions toward student learning in state-tested subject areas, and therefore, all teachers should be held accountable for student achievement growth in these areas.

Despite the rationale that all teachers should contribute to the success of students on state standardized exams, there are several disadvantages to this approach to student achievement-based evaluations. First, teachers of non-tested subjects may not necessarily teach all students in a given grade. For instance, at many schools, subjects such as music or foreign languages are electives, and are therefore not taken by all students. This raises concerns about the rigor and comparability of this evaluation model, since judgments are made about teachers based on the scores of students whom they never taught. Additionally, many teachers of non-tested subjects may view this method as unfair, since they have only a limited ability to impact student learning in tested areas. Finally, this method says relatively little about a teacher’s effectiveness in teaching their specific subject matter, and instead focuses on his or her ability to incorporate tangentially-related “tested” material into his or her lesson plans.

PRO-RATED SCHOOL/DISTRICT-WIDE VALUE-ADDED MODEL

A slightly different approach to the school/district-wide value added model is to “pro-rate” student achievement growth based on a teacher’s actual contribution to student learning in a “tested” subject area. This approach attempts to give teachers in non-tested subjects
and grades a “pro-rated” score for collaboration with teachers in tested subject areas (for instance, an art teacher who collaborates with a mathematics teacher). Additionally, some states and districts have allowed teachers in non-tested areas to choose the tested subjects used in their evaluations in order to maximize the chance that they will be evaluated on student achievement growth that most closely corresponds with their curriculum.

The drawbacks associated with this approach are similar to those noted in the broad value-added model discussed at the beginning of this section: namely that teachers of non-tested subjects have only a limited ability to affect student achievement growth in areas that they do not directly teach. An additional complication with this method is that it could be problematic and time consuming to develop metrics for accurately determining a teacher’s contribution to student learning in tested subject areas, and that differences among metrics throughout a school or state system could make it difficult to ensure comparability.

**NEW PRE- AND POST-TESTS FOR NON-TESTED SUBJECTS**

Due in part to the problems associated with basing all teacher evaluations on student achievement data for only tested subject areas, some states and districts have instead opted to create or select new standardized exams for all grades and subject areas. One obvious advantage of this model is that, when implemented successfully, all teacher evaluations can be based on student achievement growth for the specific areas in which teachers have provided instruction. Additionally, because these exams are typically created by teachers and administrators within a given educational system, they can be developed in alignment with the specific grade and/or subject standards in place within that system. Delaware is in the process of developing such exams for use in its schools as part of its Race to the Top initiative. Similarly, the Hillsborough County School District in Florida has created over 500 pre- and post- tests for 429 different courses not tested by the Florida Comprehensive Assessment Test (FCAT).³²

Writing for the National Center for the Improvement of Educational Assessment, Scott Marion and Katie Buckley note that the biggest potential drawback to this assessment method is that using tests for high stakes purposes (such as for use in teacher evaluations) “requires that the tests meet very high standards of technical quality, including a level of reliability necessary to support high stakes decisions, items and forms that meet content validity standards, and technically appropriate linking designs to ensure that scores across years or forms can be validly placed on the same scale.”³³ In order to ensure validity and comparability across years, pre- and post-tests must be re-evaluated and revised yearly—a sophisticated process which can be time consuming for teachers and administrators involved in test design.

---


Additionally, there are significant costs involved in the development, implementation, and maintenance of subject-specific pre- and post-tests. Marion and Buckley note that it is unlikely that most states or districts have the capacity to develop technically-appropriate exams “without expending significant resources for external expertise.”

**CURRICULUM-BASED END-OF-COURSE ASSESSMENTS**

Many curriculum designers include standardized end-of-course (EOC) assessments designed to measure student learning as part of their curriculum packages. Some districts and schools have considered using these existing EOC assessments as measures to assess student learning growth by administering these assessments as pre-tests at the beginning of the year and as post-tests at the end of the learning period. The advantages of this method are similar to those observed with the development and implementation of new pre- and post-tests: these EOC assessments are already well aligned with the current curriculum, and teachers for all subjects that have EOC assessments can be evaluated based on student learning growth specific to the subjects that they teach. Additionally, because EOC assessments are typically already included in curriculum packages, very few additional resources would be necessary to use these assessments for the purpose of measuring student learning growth.

The largest drawback to this approach is that curriculum-based EOC assessments are not designed to be used as pre-tests, which could potentially limit the validity of any learning growth results. Additionally, these assessments are not created and/or validated for high stakes accountability purposes, but rather to simply assess the degree to which a student has learned specific material. Therefore, it would be difficult to ensure the reliability of any results, an especially critical concern if these results are intended to be used for high stakes purposes such as teacher evaluations. Finally, curriculum-based EOC assessments do not exist for all subjects and grades, meaning that districts or states would have to either develop new pre- and post-tests for these subjects, or base teacher evaluations for these subjects on other assessment measures.

**INTERIM ASSESSMENTS**

Similar to curriculum-based EOC assessments, interim assessments can be included in curriculum or intervention packages, and are generally administered at specified intervals throughout the school year “to evaluate student knowledge and skills relative to a specific set of academic standards.” Like pre- and post-tests, these assessments can also be developed at the school or district level, provided that these educational systems possess the necessary resources for development and maintenance.

---

36 Ibid.
There are several potential advantages to using interim assessments as measures of student learning growth. First, most interim assessments are created to be aligned with state standards, and therefore should theoretically align with state standardized exams. Additionally, the National Comprehensive Center for Teacher Quality notes that schools that have already implemented response to intervention (RTI) programs have likely already adopted progress monitoring/benchmarking tools that can be used for the purposes of measuring student learning growth. These tools also have the added benefit of providing teachers with immediate feedback throughout the school year so that they can adapt their teaching methods as necessary to ensure the maximum learning growth for students in their classrooms.

Despite the fact that interim assessments are readily available and are being used by many schools and districts, there are several concerns regarding the appropriateness of these measures in determining student learning growth—particularly when these data are intended for use in high stakes teacher evaluations. Marion and Buckley note that the current interim assessment field includes many low-quality assessments, and that “their increased use could send a contradictory message about the meaning of college and career readiness as well as running the risk of narrowing the curriculum in ways that do not support the current RTTT and CCSS [Common Core State Standards] reforms.” The concern regarding college and career readiness stems from the fact that many interim exams are not particularly rigorous and may give an incorrect interpretation of standards for college and career readiness. Many researchers have also noted that these assessments may require modifications in order to ensure each interim assessment’s “validity in measuring progress on the specific content standards and its measurement reliability across students and teachers,” and that this constant evaluation of these measures may be quite costly and time consuming for districts.

**Locally-Created Assessments**

Student learning growth in some subjects—particularly those in the creative arts field—cannot adequately be measured by pencil-and-paper exams alone. Therefore, many states and districts are considering the use of “locally-created assessments” such as portfolios, performances, products, or projects (the “Four Ps”) as reliable measures of student learning growth. In order to utilize locally-created assessments effectively, teachers and administrators must first agree on a “pre-test,” administered at the beginning of the course, which is used to assess students’ existing knowledge base. Students are then asked to perform the same or similar task at the end of the course in order to demonstrate the knowledge they have gained. In many cases, the same activity can be used at the beginning and end of the course (for instance, students might be asked to play the same piece of music or perform the same monologue). For subjects for which this approach is not feasible, teachers and administrators should collaboratively “identify the

---

specific knowledge and skills that students need to know to successfully demonstrate mastery of a particular standard and then identify or develop tasks to serve as pre-tests from which progress on those standards can be determined.”

The greatest advantage of using locally-created assessments to measure student learning growth is that teachers and administrators can jointly develop assessments that accurately measure the specific content of a course. Art students, for example, would be tested on their ability to create art rather than being tested solely on their ability to recall art terms and/or apply mathematical principals to art. Additionally, this assessment method can be easily applied to any subject or grade level, making it particularly appealing to states and districts seeking assessment measures for use in teacher evaluations in non-tested areas. Both Colorado and New York are currently considering using locally-created assessments to assess teacher effectiveness in state public school systems.

However, this approach also has several potential disadvantages which may dissuade educational systems from using it for the purpose of teacher evaluations. Specifically, many researchers note that, because these assessments are typically created by teachers, concerns regarding their technical quality and comparability across classrooms may arise. Further, because performance ratings are most effective when conducted by groups of educators rather than individual teachers, extensive collaboration between teachers and administrators would be required in order to standardize assessment and evaluation rubrics across an educational system. Such collaborative efforts are generally time consuming and difficult to coordinate, and can be costly—especially if external support is employed to help with the assessment development and implementation processes. Similarly, schools may have to expend significant financial resources to train teachers on the correct administration of these assessments and evaluation of student progress.

**STUDENT LEARNING OBJECTIVES**

The Race to the Top Technical Assistance Network defines Student Learning Objectives (SLOs) as “a participatory method of setting measurable goals, or objectives, based on the specific assignment or class, such as the students taught, the subject matter taught, the baseline performance of the students, and the measurable gain in student performance during the course of instruction.” SLOs are similar in nature to locally-created assessments; indeed, any of the Four Ps can be used as assessment tools within an SLO framework. The key difference between the two assessment methods is that SLOs are typically developed at the classroom level, and are designed with the needs and capabilities of students within a specific learning environment in mind, whereas locally-created assessments are typically designed to be used across educational systems (for instance, in all fifth-grade art classes throughout a given district).

---

42 Ibid., P. 13.
44 Ibid.
In order to ensure the validity and quality of teacher-created SLOs, a rubric is often established to determine whether “the objectives and associated assessments are rigorous, measureable, reliable, and valid and whether the projected growth trajectory is considered rigorous.”47 This approach is likely popular among districts and states because teachers are encouraged to develop their own measurable goals which are in turn used in their evaluations, thereby strengthening teacher investment in the evaluation process. Additionally, as was the case with locally-created assessments, SLOs can be easily created for any subject or grade level, making them a seemingly obvious choice for states and districts seeking to link student learning growth with teacher evaluations in non-tested subject areas. Districts such as Austin Independent School District (Texas) and Charlotte-Mecklenburg Schools are already using SLOs as assessment measures in their teacher compensation plans.

Potential drawbacks of SLOs are similar to those observed with locally-created assessments. Because teachers are allowed to create their own measureable objectives, many researchers have raised concerns about the validity and comparability of these measures. Additionally, the National Comprehensive Center for Teacher Quality notes that this approach is “resource-intensive for principals or district personnel who approve objectives, provide teachers with guidance, [and] verify student outcomes.”48 Similarly, Buckley and Marion note that “this approach is only as good as the quality of the goals set for each student [or classroom] and will therefore require significant professional development in order to be able to create the learning objectives and ensure that the performance goals set are attainable yet rigorous.”49

---

48 Ibid., P. 8.
## Figure 2.1: Models of Evaluation in Non-tested Subject

<table>
<thead>
<tr>
<th>Data Category</th>
<th>Achievement Data Used in Evaluations</th>
<th>Description</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>School/District-Level</td>
<td>School/District-Wide Value-Added</td>
<td>Teachers of non-tested subjects are given the school-wide value-added average in place of individual growth results</td>
<td>No additional resources are required</td>
<td>Mathematics and reading/language arts value-added information will not be useful to teachers in improving their performance in non-tested subjects</td>
</tr>
<tr>
<td>Subject-Level Standardized Assessments</td>
<td>Pre- and Post-Tests</td>
<td>Tests that are created for areas in which few assessments exist</td>
<td>Tests can be developed in alignment with specific grade/subject standards.</td>
<td>Costly</td>
</tr>
<tr>
<td></td>
<td>Curriculum-Based End-of-Course Assessments</td>
<td>These assessments are typically included with curriculum packages</td>
<td>Tests developed by the creators of the curriculum are likely to be aligned well with the content of the course.</td>
<td>Validity is a concern whenever a measure is used in a way that was not intended by the maker of the assessment (e.g., turning end-of-course assessments into pretests).</td>
</tr>
<tr>
<td>Data Category</td>
<td>Achievement Data Used in Evaluations</td>
<td>Description</td>
<td>Advantages</td>
<td>Disadvantages</td>
</tr>
<tr>
<td>---------------</td>
<td>-------------------------------------</td>
<td>-------------</td>
<td>------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Interim Assessments</td>
<td>Evaluations are based on student performance on commercially-available interim assessment tools</td>
<td>▪ Tests are readily available for some subjects not tested by state exams</td>
<td>▪ Technical quality of current interim assessments may not be adequate</td>
<td>▪ Interim assessments do not exist for all subject areas</td>
</tr>
<tr>
<td>Locally-Created Assessments</td>
<td>The “Four Ps”—Portfolios, products, performances, or projects—are used to measure student growth over time for subjects in which standards require students to demonstrate mastery</td>
<td>▪ Evidence about student growth in particular knowledge and skills can be documented over time using performance rubrics</td>
<td>▪ Training would be required for everyone involved to ensure reliability</td>
<td>▪ Bringing raters together to examine student work may be a logistical challenge</td>
</tr>
<tr>
<td>Student Learning Objectives</td>
<td>The teacher selects objectives and determines how to assess student growth toward meeting these objectives</td>
<td>▪ Teachers benefit from being directly involved in assessing students’ knowledge and skills</td>
<td>▪ Comparability across classrooms will be problematic because of teachers’ selection of assessments and objectives</td>
<td>▪ This option is very resource-intensive for principals or district personnel who approve objectives, provide teachers with guidance, or verify outcomes</td>
</tr>
</tbody>
</table>

Source: The National Comprehensive Center for Teacher Quality

---


SECTION III: EVALUATOR TRAINING AND RELIABILITY OF EVALUATION RATINGS

Because fair and accurate observations and performance ratings depend on well-trained, consistent evaluators, evaluator training is a critical part of a teacher evaluation system. In this section Hanover discusses two central concepts in teacher evaluation: inter-rater agreement and inter-rater reliability. We then outline the evaluator training required in two pay for performance systems and the protocols that each system has established to ensure consistency among evaluators.

RELIABILITY VERSUS AGREEMENT

The U.S. Department of Education’s Center for Educator Compensation Reform (CECR) differentiates between two related concepts in teacher evaluation systems. The first, inter-rater agreement, is “the degree to which two or more evaluators give the same rating in an identical observable situations.”52 The second, inter-rater reliability, is the likelihood that two raters assign the same rankings in a set of identical situations. Because reliability is a function only of ranking and not of actual score, it is possible for raters to have a high reliability score while agreeing very infrequently. For example, two evaluators might rank teachers in the same order, but one evaluator might be consistently much harsher than the other. Because teacher evaluations typically depend on raw evaluations scores, which are unadjusted for rater severity, maximizing inter-rater agreement usually will be more desirable than maximizing inter-rater reliability. The systems profiled below, however, may sometimes use the terms interchangeably, so it is not always clear which concept is under discussion. When possible, we have attempted to maintain a clear distinction between the two concepts in our review.53

TAP

TAP teacher evaluations use a curriculum-independent set of rubrics and standards scored on a scale of one to five. A chart of these standards appears in Figure 1.1 in Section 1. Before performing observations, TAP evaluators must complete two training sessions on these standards and rubrics that involve reviews of recorded and live lessons. After training, prospective evaluators undergo a certification test using videotaped lessons; passage requires that the test-taker score each lesson on a series of indicators and deviate from the scores assigned by national raters by no more than one point (this is a test of inter-rater agreement). Subsequent to passing the rating test, the evaluator must also pass a certification test for conducting post-observation conference. Once these two

53 Ibid., pp. 1-2.
tests have been passed, the evaluator may conduct observations for a year prior to being recertified.54

TAP trains leadership teams to identify inconsistent scoring (presumably low inter-rater agreement) and to conduct scoring exercises to rectify inconsistencies. Leadership teams are expected to “ensure inter-rater reliability” by holding monthly inter-rater reliability meetings.55

**Q Comp**

Minnesota’s Q Comp system requires at least three observations of teachers during each school year. The evaluation team, which consists of two persons for each observation, must be trained and may include administrators. All evaluators must complete one day of annual comprehensive training, which includes inter-rater reliability activities. Throughout the year, instructors receive additional training to ensure inter-rater reliability.56

---

55 Ibid.
56 “Teacher Evaluation and Q Comp Comparison.” Minnesota Department of Education. p. 2.
http://education.state.mn.us/mdeprod/idcplg?IdcService=GET_FILE&dDocName=042173&RevisionSelectionMethod=latestReleased&Rendition=primary
SECTION IV: STUDENT AND TEACHER OUTCOMES

In this section we review the impact of pay for performance systems on student achievement and teacher outcomes. In order to be as comprehensive as possible, we discuss outcomes both from the programs profiled in previous sections and from additional recent initiatives. All the research is recent: no study included in this section was completed prior to 2010.

EFFECT ON STUDENT ACHIEVEMENT

To date, only one experimental study has been conducted on a large-scale pay for performance system. In this study, the National Center for Performance Initiatives (NCPI) at Vanderbilt University and RAND investigated the Nashville POINT pay for performance system during a three-year trial. The study randomly assigned middle school mathematics teachers either to a treatment group, which was eligible for bonuses tied to value-added data, or to a control group, which was not. The study's results showed that, in general, students of teachers in the treatment group did not outperform students of teachers in the control group. There was one exception to this rule: during the second and third years of the experiment, the achievement of fifth graders was positively associated with incentives to teachers. This increased achievement, however, did not persist among the same group of students after they left fifth grade.

A 2010 quasi-experimental study by Mathematica Policy Research found that TAP produced no effect on student achievement in Chicago Public schools after two years. Specifically, Mathematica detected no significant increase in performance on state tests at schools randomly assigned to the TAP program when compared to non-TAP schools. Mathematica’s follow-up study four years after the initial implementation also reported that TAP did not impact student achievement.

In 2010 a study by the NCPI found that schools in Texas’s state-funded pay for performance program, District Awards for Teacher Excellence (D.A.T.E.), had similar results on state tests to non-participating schools when controlling for the socioeconomic background of students. When this background was ignored, DATE schools had lower passing rates on state tests.

58 Ibid., p. xi.
A three-year RAND study of performance-based compensation in New York City schools discovered that the system had no impact on student achievement. In fact, state test scores at schools participating in the performance-based compensation program were slightly lower than scores at control schools, though the differences in scores were statistically significant only in mathematics and only in the study’s final year. In light of these results, RAND recommended that NYC Public Schools not resume the program.

A 2011 evaluation of ProComp performed by the University of Colorado-Denver arrived at mixed results. While Denver Public Schools students appear to have made significant achievement gains over the past several years, these gains are not clearly attributable to ProComp. Some of these gains appeared largely among student of teachers participating in ProComp, but other gains were prevalent among students of non-ProComp teachers. However, there were strong associations between earning some of the ProComp incentives and student improvement as measurement by value-added data.

**Effects on Teacher and School Culture**

In comparison to the findings on student achievement, findings on teacher and school responses to pay for performance systems are more varied. Here, we summarize some of the important findings from the studies discussed above:

- Participants in the Nashville POINT experiment thought that POINT was unlikely to distinguish good teacher from poor teachers. Over the course of the study, the percentage of teachers who believed that POINT ignored critical aspects of their performance rose from 80 percent to 85 percent. In addition, most teachers denied making any changes in response to POINT.

- Teachers participating in the D.A.T.E. program generally did not perceive negative effects from the program; however, they also doubted that the plan led to school improvements.

- Teachers participating in Denver ProComp generally agreed that the system was as fair as the traditional salary schedule. They also tended to believe that incentives for teacher evaluations were more likely to increase effectiveness than incentives related to test scores. Instituting ProComp also necessitated several beneficial operational changes in areas such as payroll, data systems, and interdepartmental communication.

---

http://www.rand.org/pubs/monographs/MG1114.html
63 Ibid., p. 29.
In NYC schools 44 percent of teachers thought that the decision-making process did not sufficiently take into account teachers’ preferences. Although most teachers expressed a desire to earn bonuses, most found the bonuses insignificant after taxes. In addition, buy-in was limited due to the belief that the evaluations relied too heavily on test scores. The pay for performance program ultimately did not alter teachers’ reports of their practices and opinions.

The two-year interim study of Chicago’s TAP program found no impact on teacher retention. However, Mathematica’s final report on TAP found increased teacher retention rates at schools that began implementing the program in 2007.

---

69 Ibid., xxv-xxvi.
CONCLUSION

In sum, recent research has uncovered a number of shortcomings and difficulties within pay for performance systems. At present researchers do not agree about the validity and reliability of the methods of data assessment that undergird most such systems. A fortiori, the appropriate weighting for student achievement data in personnel and compensation decisions remains an area of great controversy, one that cannot be settled prior to addressing the more fundamental concerns about the very use of value-added data in evaluations. Furthermore, apart from these conceptual problems, pay for performance systems may require significant expenditures for data management and additional training to ensure consistent evaluation. It does not appear that any district has yet implemented a successful pay for performance system that did not receive significant grant funding, considerably taxpayer support, or both. Most fundamentally, research has yet to demonstrate any significant impact of pay for performance systems on student achievement and teacher practice. Accordingly, it appears that a pay for performance system with widespread theoretical and empirical support from has yet to be devised.
PROJECT EVALUATION FORM

Hanover Research is committed to providing a work product that meets or exceeds member expectations. In keeping with that goal, we would like to hear your opinions regarding our reports. Feedback is critically important and serves as the strongest mechanism by which we tailor our research to your organization. When you have had a chance to evaluate this report, please take a moment to fill out the following questionnaire.


CAVEAT

The publisher and authors have used their best efforts in preparing this brief. The publisher and authors make no representations or warranties with respect to the accuracy or completeness of the contents of this brief and specifically disclaim any implied warranties of fitness for a particular purpose. There are no warranties which extend beyond the descriptions contained in this paragraph. No warranty may be created or extended by representatives of Hanover Research or its marketing materials. The accuracy and completeness of the information provided herein and the opinions stated herein are not guaranteed or warranted to produce any particular results, and the advice and strategies contained herein may not be suitable for every member. Neither the publisher nor the authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages. Moreover, Hanover Research is not engaged in rendering legal, accounting, or other professional services. Members requiring such services are advised to consult an appropriate professional.