In the following report, Hanover Research examines best practices for developing and administering teacher evaluation systems based on student growth measures. Hanover Research also examines empirical research on the validity of these systems in measuring a teacher’s contribution to their students’ academic growth.
# TABLE OF CONTENTS

**Executive Summary and Key Findings** ................................................................. 3
  - Introduction ............................................................................................................ 3
  - Key Findings ........................................................................................................ 3

**Section I: Literature Review** .............................................................................. 5
  - Overview of Student Growth measures ............................................................... 5
  - Drivers for the Adoption of Student Growth Measures ....................................... 6
  - Debates Over the Efficacy of Student Growth Measures ...................................... 7
    - Research Supporting Student Growth Measures ................................................. 8
    - Research Challenging Student Growth Measures ............................................... 9

**Section II: Best Practices** .................................................................................. 11
  - Developing Student Growth Measures ............................................................... 11
  - Selecting Participants for The Development Process ........................................... 11
  - Developing Value Added Evaluations .................................................................. 11
  - Selecting Teacher and Student Assessments for Evaluations ............................ 13
  - Administering Student Growth Measures ........................................................... 17
    - Explaining the System to Stakeholders .............................................................. 17
    - Evaluating Teachers in Student Growth Systems .............................................. 18
  - Collecting and Analyzing Data for Teacher Evaluations .................................... 19
    - Planning for Data Collection and Analysis ......................................................... 19
    - Data Parameters for the Collection Process ....................................................... 20
    - Gathering Student Demographic Information .................................................... 21
    - Carrying Out Direct Observations .................................................................... 21
    - Technical Components Related To Gathering Data .......................................... 22
EXECUTIVE SUMMARY AND KEY FINDINGS

INTRODUCTION

This report examines best practices and empirical literature on teacher evaluation systems based on student growth measures. In recent years, these systems have become key components of state and school district evaluation programs across the United States. Spurred by federal initiatives such as the 2009 Race to the Top (RTTT) program, student-growth teacher evaluation systems such as value-added models assume that evaluators can isolate and measure a teacher’s contribution to their students’ academic growth. While the validity of these models have generated contentious debates among researchers, school districts and experts have sought to develop best practices for improving the capacity of these systems to improve teacher effectiveness and student learning.

This report comprises two sections to examine these two topics:

- **Section I: Literature Review** will examine empirical research on these systems, including both studies that support and challenge the validity of measuring a teacher’s contribution to student growth.
- **Section II: Best Practices** will examine best practices for developing and administering these systems as well as gathering evaluation data for analysis.

KEY FINDINGS

- **Reviews of empirical research studies on the validity of student growth measures – especially value-added measures – do not provide definitive support for the adoption of these systems.** In particular, these findings suggest that the lack of common models and rubrics for different student growth measures makes it difficult to draw conclusions about their validity in measuring teacher effectiveness. This literature also highlights several problems related to error rates, inaccurate test effects, and model biases that can undermine the integrity of personnel decisions based on the distortions in these ratings.

- **While individual empirical studies do support the validity of student growth measures, some of these studies provide critical qualifications that districts should consider when adopting them.** In particular, these studies recommend that evaluators should incorporate observational data and acknowledge the influence of non-test factors when reviewing teacher evaluations in these systems.

- **Value-added systems should utilize either standard versions of the model or a two-step model that de-links teacher-level measures of test-score growth and student characteristics.** In particular, two-step models can produce consistent results that can generate teacher effectiveness data for improving teaching performance. Districts should also use absolute teacher evaluation scores and ranking scales with interpretable cut points to evaluate teachers in these systems.
School districts should not utilize value-added measures to make high-stakes decisions such as firing teachers or awarding tenure status to instructors. According to best practices experts, schools and districts may make disciplining errors based on the results of poorly-designed value-added assessment models. Instead, experts suggest that school districts should use these models to inform formative assessments, guide professional development, and provide feedback to teachers on their performance.

The teacher evaluation process in student growth systems should not consist of a single rating assigned at the end of the year. Instead, the evaluation process should include frequent classroom observations followed by timely, constructive feedback from the evaluator. Evaluators and teachers should also have frequent conversations about teachers’ classroom performance to create support measures that address their needs in promoting student growth.

Data collected through direct and indirect teacher observations should form an important input into student growth models – especially for value-added models. Empirical research and best practices literature state that this information complements the ratings in these systems by providing better insight into teacher effectiveness based on a teacher’s actual educational practice to increase student learning. Experts recommend that principals, teacher peers, and third-party evaluators should carry out these observations.
SECTION I: LITERATURE REVIEW

This section will examine empirical literature on student growth measures, with a focus on value-added measures and student growth percentiles for teacher evaluations. The first section will provide an overview of student growth models, including value-added measures and student growth percentile models. The second section will discuss the federal policies driving the adoption of these forms of teacher evaluations across U.S. states and school districts. Finally, the last section will examine empirical debates over the validity of these evaluation methodologies, including a discussion of research that supports or challenges the value-added method and other forms of student growth measures.

OVERVIEW OF STUDENT GROWTH MEASURES

Broadly, teacher evaluations that use student growth measures isolate and measure a teacher’s contributions to their students’ academic growth. These models, which use different estimation methodology and control variables to calculate teacher effectiveness, fall into five categories. Figure 1.1 shows the different models and their key components.

Figure 1.1: Student Growth Teacher Evaluation Models

<table>
<thead>
<tr>
<th>MODEL TYPE</th>
<th>PURPOSE</th>
<th>CALCULATES</th>
<th>INTENDED USE</th>
<th>EXAMPLES OF STATES IN USE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trajectory</td>
<td>Uses a trajectory where students must close an achievement gap in three to four years</td>
<td>Actual or projected growth</td>
<td>Rewards growth that meets trajectory</td>
<td>Florida</td>
</tr>
<tr>
<td>Projection</td>
<td>Estimates an individual’s future performance based on past scores</td>
<td>Actual or projected growth</td>
<td>Rewards growth</td>
<td>Arkansas, Iowa, Ohio, Pennsylvania</td>
</tr>
<tr>
<td>Value Table</td>
<td>Assigns points based on yearly improvement or maintenance of proficiency</td>
<td>Closing proficiency gap or proficiency maintenance</td>
<td>Rewards growth</td>
<td>Arkansas, Delaware, Iowa</td>
</tr>
<tr>
<td>Value-Added</td>
<td>Rates schools based on changes in student performance that are better than expected</td>
<td>Better than expected growth</td>
<td>Rewards better than expected growth</td>
<td>Florida, Ohio, Pennsylvania</td>
</tr>
<tr>
<td>Student Growth Percentiles/Colorado Model</td>
<td>Provides information about an individual or group’s progress to parents, teachers, and administrators</td>
<td>A student or group’s relative growth</td>
<td>Rewards growth</td>
<td>Colorado, Arizona</td>
</tr>
</tbody>
</table>

Source: The Statewide Longitudinal Data Systems (SLDS) Grant Program

Among these models, the value-added and “Colorado” models have emerged as key approaches for evaluating teacher effectiveness. Value-added models measure teacher or school effects on a student’s learning while separating the effects of non-academic factors on student growth, such as family and peer influence. In this manner, the model seeks to see “the value added” that a teacher or school contributes to their students’ growth. On a technical level, this model uses an equal interval scale such as a Normal Curve Equivalent score to measure a teacher’s effects on student growth between two points. Evaluators treat value-added performance ratings as a causal relationship between teacher effectiveness and student learning since they use student covariates.

In the case of the student percentile growth model, or the “Colorado model,” the model compares a student’s progress to other students with historically similar scores and uses these comparisons to predict future performance through quartile regressions. In particular, the model compares the student’s growth percentile to similar students, which educators interpret as the percentage of the student’s academic peers with lower scores in the last year. For instance, a student growth percentile of 33 means that 33 percent of the student’s academic peers received lower scores on the most recent test. In contrast to value-added models, this model serves as descriptive measures of growth rather than causal measures since it does not use covariates in the calculations.

DRIVERS FOR THE ADOPTION OF STUDENT GROWTH MEASURES

Although states such as Tennessee and districts such as Denver Public Schools have utilized student growth measures since the 1990s, two federal initiatives have driven the expansion of these methods across the United States. In 2009, the Department of Education adopted the Race to the Top program (RTTT) and Teacher Inventive Fund, which offered grants to states and school districts that adopted teacher evaluation measures as a part of broader school reform strategies. As a part of this measure, the program encouraged states and districts to link teacher contribution to student growth to compensation, renewal, and tenure. In 2011, the U.S. Department of Education’s ESEA Flexibility program began allowing states to avoid federal sanctions if they adopted teacher evaluation programs based on student growth measures.

The incentive for these federal programs stems from the weaknesses of many teacher evaluation systems in states and school districts. For example, a 2009 study by Weisburg et al. finds that districts that use binary ratings gave over 99 percent of teachers a satisfactory rating and 95 percent of teachers received one of the top two ratings in districts with more

---

7 Ibid., p. 4.
rating categories. These assessments are problematic because they obscure the fact that teacher effectiveness varies from individual to individual; these ratings also fail to incorporate assessments by administrators or other teachers. As a result, these systems fail to provide an accurate measure of a teacher’s effectiveness in promoting student learning.

In response to these federal initiatives and research identifying the weaknesses of traditional forms of teacher evaluations, states such as Tennessee and Colorado have adopted state-wide teacher evaluation measures that incorporate student growth assessments or allow local school districts to work with teachers to devise and administer their own classroom assessments. States have also increased the number of tested subjects and grades to ensure that more teachers can participate in these evaluations, and are utilizing different methods to measure teacher contributions to student growth in non-tested subjects.

DEBATES OVER THE EFFICACY OF STUDENT GROWTH MEASURES

Although many studies agree that teachers impact student achievement, much of the literature on student growth measures addresses ongoing debates about their validity and accuracy in evaluating a teacher’s contributions to a student’s growth. In addition to promoting the validity of these measures, student-based growth advocates offer the following arguments in favor of adopting these measures:

- Merit-based payment will incentivize teachers to improve student learning, generating better test scores.
- Teachers that are more effective than others deserve to have better compensation.
- The education profession can attract and retain higher-quality candidates and teachers if compensation schemes reward high-quality teaching.

However, critics of these models argue that these approaches – especially value-added models – do not provide accurate measures of teacher impact on student learning due to problems such as data distortions and understating the role that the environment plays on student outcomes. As the President of Math for America argues, the press has promoted value-added models without acknowledging the body of academic research that highlights the limitations of these systems, especially in the use of high-stake decisions such as hiring and firing teachers.

———

10 Ibid.
The next two subsections will examine the empirical research on the validity of these models. This literature includes journal publications and scholarly studies produced after 2011 to ensure relevance to current findings or beliefs. While the review is not a definitive summary of empirical research on these models, it demonstrates the large number of studies that question the validity of these approaches.

**Research Supporting Student Growth Measures**

Among the literature that supports the validity of student growth models, recent individual studies find that these methods can evaluate teacher effectiveness and the development of student learning. However, some of the examined studies qualify their findings by noting that many of the methodologies have limitations in data that can impact the accuracy in the measurement of teacher effectiveness.

One 2011 study published in *Educational Assessment, Evaluation and Accountability* examined the validity of value-added measures that used observation data, and concluded that value-added scores can “give teachers an indication of student growth within their classroom and provide incentive to examine curriculum and instruction if these scores are used as a guiding tool rather than as a high-stakes gavel.” However, the study adds that these models rely on teacher observation data to identify ineffective teachers, teacher improvement areas, and different regions of a district that need effective teachers. In the author’s opinion, these data would improve the interpretation of value-added analyses while minimizing the misinterpretation of student growth data or the misclassification of teachers’ effectiveness in low-income schools.16

Further, a 2012 study in *Education Finance and Policy* used simulations to determine the ability of various value-added approaches to produce accurate estimates of teacher effects on student achievement. The study found that teacher rankings in these systems “can correlate relatively well with true rankings in certain scenarios.”17 The study also found that misclassification rates may be relatively lower than opponents of these measures acknowledge. However, the study notes that the impact of non-test factors, such as the context of an instructor’s classroom, on value-added measures means that evaluators must use close scrutiny “regarding the [evaluation] methods used and interpreted with a high degree of caution.”18

Finally, a 2013 U.S. Department of Education-issued literature review on alternative growth measures that form a major component of teacher evaluation systems suggests that “locally developed, curriculum-based assessments or non-test outcomes” may serve as reliable differentiators among teachers and schools. The literature review notes that two studies

18 Ibid.
that examined the use of these assessments found that they can reliably distinguish teachers at different points of their scale’s performance range. However, the studies did not analyze outcomes for subjects without tests for individual teacher performance.\textsuperscript{19}

**RESEARCH CHALLENGING STUDENT GROWTH MEASURES**

Among studies that do not support the validity of student growth measures, cross-research reviews have not found definitive evidence that these systems serve as accurate measurements of teacher effectiveness, especially for value-added models. More importantly, this literature highlights several problems related to error rates, distorted test effects, and model biases that appear in these evaluations.

First, researchers have identified key problems related to the application of these systems. In a 2013 review of available research on value-added measures in education, Gabriel and Lester note three key problems with this model: error rates, distorted test effects, model biases. In the case of error rates, the authors note that scholars have highlighted how error rate problems related with value-added ratings can lead to poor decisions for high-stakes decision-making, such as hiring or firing teachers.\textsuperscript{20} The authors also note that studies such as the Gates Foundation’s Measures of Effective Teaching Project (MET) find that value-added scores from state tests and open-ended tests have low correlations for English and math performance. Finally, the authors highlight studies such as Chetty, Friedman, and Rockoff, (2012) and Kane and Staiger (2008) that demonstrate how different variables can generate model bias that impacts the results and stability of teacher evaluations.

Cross-literature research has also found that the many empirical studies of these systems do not provide sufficient evidence for their validity. In a 2013 review of literature on value-added systems, Murphy et al. argue “that the evidence for using VAMs [value-added models] to evaluate teaching is insufficient due to the inconsistent and overstated magnitude of the effects of teaching-related variables on student learning outcomes and gains.”\textsuperscript{21} In another review of literature on the impact of these models on student learning, Spooren, et. al. (2013) note the lack of “consensus on the degree of association between student ratings and learning outcome measures.”\textsuperscript{22} The researchers also found considerable variation in the coefficients of the models in their literature review, suggesting “that there are many other factors that account for learning” of their students.\textsuperscript{23}

In another 2013 summary of empirical literature on the validity of value-added models, Graue, Delaney, and Karch note similar trends in empirical findings related to the reliability and validity of these models stemming from their overreliance on standardized tests, use of


\textsuperscript{22} Ibid.

\textsuperscript{23} Ibid.
compromised data, and struggles with biases that compromise their modeling of teacher effectiveness.24 These authors also note problems related to the applications of value-added models to education systems, including issues related to inconsistent results where a teacher adds value in one model while failing to generate results in others. The authors also note studies that highlight the lack of stability in teacher scores over time.25

In a 2013 cross-research study exploring how different data and model configurations impact value-added scores, Kersting, Chen, and Stigler note that the vast number of configurations for value-added measures makes it difficult to establish which models and data configurations are the most effective measures of teacher effectiveness. As they argue,

Understanding which data and model specifications have a larger effect on teacher VAEs [value-added evaluations] compared to others [evaluation models] has been challenging because the number of model specifications and their possible combinations is too large to be fully investigated in any single study.26

The authors subsequently note that only one study – Papay (2011) – uses a systematic approach to explore the effects of different variables on value-added scores. Instead, most studies examine only a small number of value-added approaches for these validity studies.27

More critically, the authors also note that many of these studies do not use common metrics to report on the stability of teacher value-added scores over the course of the implementation of these evaluation systems. As they note, different studies divide the value-added distribution into different categories of performance groups. In some instances, some studies use three groups, such as “average effective,” “significantly above average effective,” and “significantly below average effective,” while other studies use four or five groups to compare teacher ratings. As a result, the authors argue that these inconsistencies in categories makes it difficult for researchers and districts to compare percentage changes in teacher designations and derive meaningful understandings of the validity of value-added approaches to teacher evaluations.28

In response to these problems related to the inconclusive nature of these studies, the same authors advocate for incorporating teacher behavior and practice observations into these evaluation metrics to determine teacher effectiveness. As they note, “observational evaluations can yield a deeper understanding of teacher effectiveness outcomes based on what a teacher does in the classroom to increase student learning.”29 As the next section will demonstrate, teacher observations are a key best practice for developing effective student growth evaluation systems for teachers.

25 Ibid.
27 Ibid.
28 Ibid.
29 Ibid.
SECTION II: BEST PRACTICES

This section will present best practices for developing and administering student growth measures, as well as for gathering data for teacher evaluations. Information in this section is drawn from empirical research, industry publications, and handbooks for individual school districts. The first section will examine different components of developing student growth measures, including selecting evaluation categories and rubrics for value added measures. The second section will focus on administering these measures, including providing teachers with feedback during the evaluation process. The last section will discuss the process of selecting and collecting data for teacher evaluations, including issues related to incorporating student demographic information into teacher evaluations.

DEVELOPING STUDENT GROWTH MEASURES

SELECTING PARTICIPANTS FOR THE DEVELOPMENT PROCESS

During the initial phases of developing student growth measures, districts should organize groups that include input from teachers and teachers unions to develop their student growth measures. Denver Public Schools, for instance, gathered district staff and the Denver Classroom Teachers Association (DCTA) to design the city’s teacher evaluation system using ideas from teachers and administrators. Tulsa Public Schools also incorporated teachers and teacher feedback into the process of developing its teacher evaluation system.

However, some states give district superintendents greater control over this process. In Connecticut, superintendents largely develop and implement teacher evaluation systems and processes for their districts. While these administrators must follow state guidelines, they have some flexibility in creating an evaluation system that meets their district’s needs as long as the process aligns with the state’s department of education guidelines and agreements over evaluations that exist between local boards of education and teacher unions.

DEVELOPING VALUE ADDED EVALUATIONS

SELECTING VALUE ADDED EVALUATIONS

Best practices literature on creating value-added evaluations recommends two specific models for evaluating teachers: a standard value-added model and a two-step model. In the

case of the former, the National Education Association (NEA) recommends that districts and states use a standard value-added model that includes prior information about student achievement and student, classroom, and school characteristics such as a student’s socioeconomic status. While the NEA believes that this model will not generate biased estimates of teacher effectiveness compared to other models such as growth or attainment evaluation, it strongly advises districts to avoid using this model for high-stakes decision due to “significant limitations that make them inappropriate options for these situations.”

In contrast, research published in a 2014 Education Next article suggests that districts should use a two-step value-added model that delinks school- and teacher-level measures of test-score growth and student characteristics. As the researchers argue, this method is preferable to other growth models because it allows districts to ensure “that the comparisons used to measure performance are between schools and teachers that are in similar circumstances.” In the process, districts can develop effective evaluations for teachers and schools that generate performance data for improving instruction. This approach also avoids compounding inequities between schools serving advantaged and disadvantaged students.

**DEVELOPING SCORING RUBRICS FOR VALUE ADDED EVALUATIONS**

Broadly, researchers and experts recommend that districts should use a combination of absolute teacher evaluation scores and ranking scales with interpretable and fair cut points to evaluate teachers. The research recommends using absolute measures, which score teachers independently of their peers, over relative measures, which scores teachers relative to their peers, as relative scores can obscure teacher contributions by ranking teachers from highest to lowest rather than determining whether an individual teacher met their target level of effectiveness.

Two examples highlight the advantages of absolute reporting system over a relative one. In a district with high performing teachers that uses relative reporting systems, the bottom 5 percent of instructors may have made significant contributions to student learning even though they appeared in the bottom tier of their peer rankings. In contrast, an absolute system tells a teacher that they received a score of 200 on a scale that runs from 100 to 250, which allows them to understand where they can improve their teaching practices.

In order to improve an absolute reporting system, districts can assign cut points to these scales for different levels of performance. On a scale from 100 to 250, for example, 100 to 149 might be defined as “low-performing,” 150 to 199 is “satisfactory performance,” and

---


35 Ibid.


37 Ibid.
200 to 250 is “high-performing.”38 Districts should ensure that the design of the reporting scale is transparent and includes input from a variety of stakeholders, including unions.39

The Brookings Institution also recommends that districts should eliminate or reduce to a low weight the contribution of school-wide value added scores to individual teacher evaluations. As the researchers argue, the inclusion of this score, which stems from efforts to promote teacher cooperation, negatively impacts “good teachers in bad schools and positively impacts bad teachers in good schools.”40 If districts want to develop strategies to promote greater cooperation, the Brookings Institution suggests that they can consider modest school-based bonuses or evaluation of individual teachers based on their contributions to activities that benefit learning outside of the classrooms.41

SELECTING TEACHER AND STUDENT ASSESSMENTS FOR EVALUATIONS

SELECTING TEACHER PERFORMANCE ASSESSMENTS

One of key components of student-growth measures revolves around selecting the assessments for evaluating teacher performance. Figure 2.1 below outlines an example of the guidelines for best practices in student growth assessments, as promoted by one Connecticut-based child-rights coalition.

Figure 2.1: Recommendations for Developing and Selecting Student Growth and Teacher Assessments

<table>
<thead>
<tr>
<th>EXPECTATION</th>
<th>COALITION’S DESCRIPTION OF THE EXPECTATIONS</th>
</tr>
</thead>
</table>
| Expectations Should Be Clear and Rigorous | • Teachers should be evaluated against clear, rigorous performance expectations that reflect excellence in classroom teaching and promote student learning.  
  • Performance expectations should be precise and clear to promote consistency in the evaluation process. |
| Expectations Should Focus on Specific, Observable Student Behavior | • Districts should find evidence that students are actively engaged in the lesson being taught.  
  • Evaluators need a clear, workable assessment and scoring tool to help them make consistent judgments when evaluating teachers against performance expectations. |

38 Ibid.
39 Ibid.
41 Ibid., p. 14.
<table>
<thead>
<tr>
<th>EXPECTATION</th>
<th>COALITION’S DESCRIPTION OF THE EXPECTATIONS</th>
</tr>
</thead>
</table>
| Expectations Should Use Multiple Measures | ▪ Evaluation systems should use multiple measures to determine whether teachers meet performance expectations.  
▪ These measures include classroom observations, as well as objective evidence of student academic improvement.  
▪ Some examples include student performance on standardized tests and district- and teacher-generated assessments.  
▪ Each measure should have a specific weight so that the teacher and evaluator know how much effect each measure will have on the teacher’s overall evaluation rating. |
| Expectations Should Use Multiple Ratings | ▪ Each measure of teacher performance being evaluated should be assigned one of four to five rating levels to give teachers a clear picture of distinct differences in their performance.  
▪ For example, a five-tier rating system might use: 5 = Exemplary, 4 = Strong, 3 = Effective, 2 = Developing, 1 = Needs Improvement |

Source: The Bridgeport Child Advocacy Coalition

The National Comprehensive Center for Teacher Quality also recommends that teacher assessments should aim to:
▪ Include protocols and processes that teachers can examine and comprehend;  
▪ Directly align with teaching standards;  
▪ Motivate teachers to examine their own practice against specific standards;  
▪ Allow teachers to participate in or co-construct the evaluation;  
▪ Give teachers opportunities to discuss the results for formative purposes with evaluators, administrators, teacher learning communities, mentors, and coaches; and  
▪ Inform professional growth and development offerings.

At the district level, Tulsa Public Schools provides an example in using best practices research to ensure the system assesses a teacher’s ability to fulfill the three categories of responsibilities that promote student growth and achievement. Figure 2.2 on the following page outlines these categories, including the “Cotton paper” category that uses recommendations from a report written by Kathleen Cotton for the Association for Supervision and Curriculum Development, and the “Kane Paper” category that lists recommendations from a paper written by Harvard Graduate School of Education economist Thomas Kane.

---

Figure 2.2: Tulsa Public Schools Teacher Evaluation Categories

<table>
<thead>
<tr>
<th>GROUP</th>
<th>TULSA PUBLIC SCHOOL DESCRIPTION OF BEST PRACTICES</th>
</tr>
</thead>
</table>
| “Cotton Paper” Best Practices | ▪ Clearly communicate and support high behavioral expectations.  
▪ Consistently apply rules and standards of behavior.  
▪ Stop disruptions quickly and maximize learning time.  
▪ Differentiate and adapt instruction to the needs of faster and slower learners.  
▪ Pace lessons appropriately, including minimizing transition time and monitoring student progress. |
| Increasing Student Achievement | ▪ Explain lessons and objectives clearly.  
▪ Describe the relationship of the current lesson to previous learning.  
▪ Use strategies such as advance organizers.  
▪ Ask questions that engage student interaction and enable the teacher to monitor student understanding.  
▪ Provide for “wait time” when questioning students.  
▪ Give timely feedback and reinforcement. |
| “Kane Paper” Best Practices | ▪ Clearly define and support expected behavior.  
▪ Develop plans to achieve identified objectives.  
▪ Use higher-level questioning techniques.  
▪ Engage all learners.  
▪ Differentiate instruction to respond to different student needs.  
▪ Provide adequate and timely feedback.  
▪ Adjust instruction based on the results of monitoring.  
▪ Create a caring, respectful and effective learning environment. |

Source: Tulsa Public Schools

Two empirical studies have proven that these categories are effective measures of teacher impact on student performance. A study carried out by the Bill and Melinda Gate’s Foundations MET Validation Engine study finds that the district’s teacher evaluation captures practices that correspond with gains in student achievement. 45 A research team from the University of Wisconsin’s Value-Added Research Center finds similar results: the teachers’ overall evaluation scores across these categories correlated with their value-added scores and student growth measures. 46

In Denver Public Schools, the district utilizes two broad categories for evaluating a teacher: student growth and professional practice. Figure 2.3 on the following page displays the components of these categories.

---

46 Ibid., p. 5.
The district allocates 50 percent of the score to student growth and 50 percent to professional practice, which focuses on observation scores and ratings. In the case of professionalism, the district defines this category as the individual and collective work teachers carry outside the classroom to support students’ learning. Student perceptions are gathered through student surveys that measure a teacher’s ability to facilitate learning, support students, and communicate high expectations.

A similar model is used in Cincinnati Public Schools, where the district introduced an evaluation model that links student progress to pay raises and advancement. In the model, one half of the evaluation score uses classroom observations by a certified consulting teacher or trained teacher evaluator while the other half uses student growth scores to measure teacher effectiveness.

**SELECTING STUDENT GROWTH ASSESSMENTS**

Another key component of developing student growth systems revolves around selecting the appropriate assessments for measuring student growth over time. Broadly, student growth models should use assessments that align with state or district curriculum. When these tests and curricula do not align, models such as value-added measures will provide inaccurate information about a teacher’s performance and may even penalize teachers who

---

48 Ibid., p. 7.
follow the curriculum rather than “teaching for the test.” In many instances, teacher ratings such as value-added scores use standardized student achievement tests and attempt to measure the amount an individual teacher contributes to student learning.  

At the district level, Denver Public Schools uses the following student growth assessments for its teacher evaluation programs. It should be noted that the district includes school- and district-level measures of student growth in teacher evaluations, although the benefits and drawbacks of this model are questioned by experts and reviewed in the previous subsection of this report.

<table>
<thead>
<tr>
<th>ASSESSMENT CATEGORY</th>
<th>DISTRICT’S DESCRIPTION OF THESE MEASURES</th>
</tr>
</thead>
</table>
| State Measures       | • Measures the growth of a teacher’s own students on state tests.  
                       | • This category applies to teachers who instruct grades 4 to 10 in the state-tested subject areas of reading, writing, and math. |
| School Measures      | • Measures academic growth on the district’s School Performance Framework (SPF).  
                       |   o This evaluation applies to all teachers who received teaching positions in the last year.  
                       | • Measures progress toward students’ learning goals using interim assessments, performance tasks, and unit assessments. |
| District Measures    | • Measures growth on the state’s District Performance Framework (DPF).  
                       | • Also measures student growth on state tests and ACCESS. |

Source: Denver Public Schools

The National Comprehensive Center for Teacher Quality provides similar recommendations. The organization suggests that districts use standardized tests, pre or post-course tests in untested subjects, student performance for visual and performing arts, curriculum-based tests that used standardized formats, and classroom-based tests such as DIBELS.

**ADMINISTERING STUDENT GROWTH MEASURES**

**EXPLAINING THE SYSTEM TO STAKEHOLDERS**

One of the key components of administering student growth administration is explaining the system and its components to stakeholders such as teachers and parents. Given that these evaluation systems are complex and diverge from traditional methods for reporting student learning and evaluating teachers, districts should produce accessible materials that explain:

- Why the district is adopting this system for teacher evaluations;
- How these evaluations differ from previous evaluation systems;

---

- How the district with administer the evaluations; and
- The new system’s implications for students, parents, and teachers, including compensation and evaluations for teachers.\(^{54}\)

The district should prepare these materials with input and feedback from key stakeholders, including teachers and the teachers’ union.\(^{55}\)

**EVALUATING TEACHERS IN STUDENT GROWTH SYSTEMS**

In terms of best practices for the utilization of evaluation ratings, *researchers advise against the use of value-added ratings for high stakes decisions such as dismissals, promotions, or compensation*, since the potential inaccuracy of these evaluations that stem from poor data or poorly designed systems could obscure a teacher’s true effectiveness.\(^{56}\) Instead, value-added scores should inform formative assessments, guide professional development, and provide feedback to teachers.\(^{57}\)

Researchers recommend that the teacher evaluation process should not consist of a single rating assigned at the end of the year. Instead, the evaluation process should include frequent classroom observations followed by timely, constructive feedback from the evaluator. Evaluators and teachers must have frequent conversations about their classroom performance, student progress, professional goals, and developmental needs, and create support measures that address these needs.\(^{58}\) At the end of these conversations, teachers and evaluators should reach an understanding of the teacher’s short-term goals and the manner in which the evaluator and other school staff will help them improve performance.\(^{59}\)

At the district level, Denver Public Schools utilizes different approaches for evaluating and providing feedback to teachers based on the two major categories of evaluation: student growth and professional practice. Figure 2.5 displays the steps that the district recommends that teachers and evaluators use to evaluate teacher performance.

---


\(^{55}\) Ibid.

\(^{56}\) Ibid.


\(^{59}\) Ibid.
In addition to these approaches, experts also recommend that districts should release individual teacher scores along with detailed technical documentation that allows value added or evaluation scores to be replicated. The district can provide examples of the content of the technical documentation and allow for external evaluation.61

COLLECTING AND ANALYZING DATA FOR TEACHER EVALUATIONS

PLANNING FOR DATA COLLECTION AND ANALYSIS

In terms of allotting time for analyzing data gathered from evaluations, districts should create reasonable estimates for the amount of time required to gather and analyze data based on state and district testing schedules and the amount of data to be analyzed. The timeline should allow sufficient time for data collection, verification, correction, analysis, and reporting. The timeline should also meet state and district needs, including allowing time for the district to identify low performers for professional development or recruit high performers to work in priority schools.62

An example of this time frame from Denver Public Schools is presented in Figure 2.6 on the following page. The DPS timeline incorporates the above recommendations into five steps.

62 Ibid., p. 22.
DATA PARAMETERS FOR THE COLLECTION PROCESS

Districts should gather data from classes with a minimum of 15 students per teacher for each year in the evaluation. In particular, the literature states that larger classrooms with 20 or more students are preferable since smaller class sizes introduce more imprecision in isolating a teacher’s contribution to student learning. As noted by Baer, et. al. (2010), “generally, the more students in a classroom, the greater the precision of the VAM [value-added model] score for a teacher.” In the case of student absenteeism and student mobility, districts can collect daily student attendance records and use this information to adjust teacher evaluation scores. Additionally, data systems should allow evaluators to track student growth across schools in the same district.

Districts should use a three-year rolling average when calculating a teacher’s evaluation scores for a year. Specifically, administrators and evaluators should calculate the score for three consecutive years and calculate the average to generate the effectiveness score. This score for one year represents the average of a teacher’s scores over the last three years in that grade and that subject. This time frame allows for variation in actual teacher effectiveness and addresses the imprecision of evaluations such as value-added measures.

---

65 Ibid., p. 13.
66 Ibid., p. 9.
GATHERING STUDENT DEMOGRAPHIC INFORMATION

A review of best practices literature found a divide over whether student growth systems should use student demographic data, especially for value-added modeling. Among advocates for including this information, a Brookings Institution report recommends that evaluations such as classroom observations should incorporate adjustments for student demographics. The researchers’ analysis of data in four districts across the United States finds that statistical adjustment of classroom observation scores for student demographics is successful in producing a pattern of teacher ratings that approaches independence between observation scores and the incoming achievement level of students.67

Baer, et. al. (2010) also suggest that districts should include student demographic information and additional background characteristics for student motivation, parental support at home, and peer influences when generating VAM scores that impact student achievement. These data may allow districts to produce more precise estimates of what the teacher and school – aside from home and community – contribute to student learning.68 School districts in New York and Washington DC with value added models use demographic information such as race, gender, and socioeconomic class for teacher evaluations.69

However, other researchers argue that adjusting for student demographic characteristics is unnecessary, as the growth scores are calculated by comparing students against themselves. In their opinion, including student characteristics could distort the scores by making it easier for teachers of disadvantaged students to earn higher scores. For example, this model would lower the expectations for black student – and their teacher’s performance – based on the assumption that black students do not perform at the same achievement level as white students.70 Ultimately, districts must use the merits of these two positions to determine whether incorporating student demographic information is necessary for evaluating teacher effectiveness.

CARRYING OUT DIRECT OBSERVATIONS

As a part of administering teacher evaluations, many districts will incorporate data gathered from direct observations of teacher as a part of calculating a teacher’s rating. In order to maximize the efficacy of these observations, researchers and experts recommend that districts should conduct two-to-three annual classroom observations for each teacher; at least one of these observations should involve a trained observer from outside the school without substantial links to the teacher. Districts should arrange for an additional classroom observation by another independent observer when substantial differences exist between a teacher’s observation scores.71

70 Ibid.
School districts have also developed best practices for teacher observations. In the case of Denver Public Schools, the district reworked its observation processes in 2011 to improve its capacity to coach and support teachers. The district based a part of the teacher evaluation model on four annual classroom observations, two conducted by their principal and two by their peers. The district also developed a set of detailed standards to guide observations and a set of videos to illustrate what effective practices might look like in the classroom.\(^{72}\) Research on a pilot observation program in Chicago suggests that scheduled and unscheduled observations present a complete picture of a teacher’s actual effectiveness.\(^{73}\)

**TECHNICAL COMPONENTS RELATED TO GATHERING DATA**

Districts should have the technological infrastructure to link teachers to students and address problems related to missing student data. These measures are important since data errors can threaten the credibility of the evaluation system, especially if they produce errors in bonus calculations or discredit teachers based on students they did not teach.\(^{74}\)

Districts should also have data systems in place that link teachers, students, and subject-specific test results from year to year. This data infrastructure is importance because must be able to link teachers, their students, and results from subject-specific student test scores to generate evaluation scores. Moreover, districts must have the capacity to track students over time and from school to school within a district.\(^{75}\)

Districts must also have measures in place to address problems related to missing student data, including plans for adjusting scores that use recoverable and non-recoverable missing data. In the case of recoverable data, districts should modify data systems to ensure that this data is included in the score calculation. In the case of non-recoverable missing data, districts should measures in place to minimize bias in score calculation.\(^{76}\)

---


\(^{75}\) Ibid., p. 11.

\(^{76}\) Ibid., p. 12.
PROJECT EVALUATION FORM

Hanover Research is committed to providing a work product that meets or exceeds partner 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 that 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 partner. 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. Partners requiring such services are advised to consult an appropriate professional.