

## 4.0 - SITE VISIT REPORT – FINDINGS & RECOMMENDATIONS

### 4.1 Case Study Methodology

The case studies examined the relevance and usefulness of the OTES model for guiding LEAs' implementation of high-quality teacher evaluation methods aligned to the Ohio Standards for the Teaching Profession and best practices in measuring teacher quality. Between March 26 and April 5, 2012, two evaluators from MGT of America conducted site visits at 12 LEAs that were selected from a stratified random sample of OTES pilot sites between March 26 and April 5, 2012. The types of districts and geographic regions in the statewide distribution of case study sites are explained in *Exhibit 1*.

EXHIBIT 1  
TYPE AND GEOGRAPHIC REGION OF RANDOMLY SELECTED LEAs

OTES with Student Growth Measures					
Case Study Site 1	Case Study Site 2*	Case Study Site 3	Case Study Site 4	Case Study Site 5	Case Study Site 6
Exempt Villages	Cities	Locals	Locals	Locals	Exempt Villages
Northeast	Urban	Northwest	Central	Northeast	Northwest

OTES without Student Growth Measures					
Case Study Site 1*	Case Study Site 2	Case Study Site 3	Case Study Site 4	Case Study Site 5	Case Study Site 6
Cities	Cities	Education Service Center	Cities	Locals	Community School
Southwest	Urban	Northeast	Northwest	Southeast	Southeast

\*These two sites included locally developed evaluation system components in their OTES pilot

MGT evaluators conducted a total of 75 interviews during the site visits. The percent of interviews for each role is identified in *Exhibit 2*. Of the interviews, 42 percent were LEA superintendents and principals, 36 percent were OTES pilot teachers, and 21 percent were non-participating principals and teachers.

EXHIBIT 2  
PERCENT OF LEA INTERVIEWEES BY ROLE

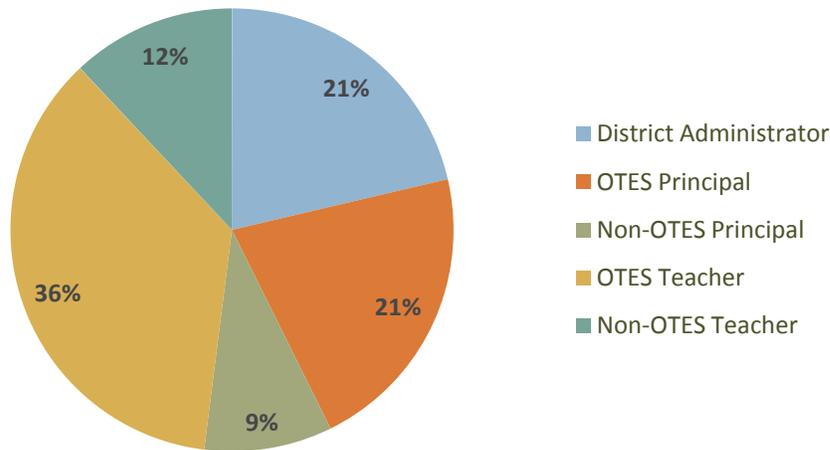


Exhibit 3 shows that at 92 percent of the cast study sites, evaluators were able to interview the LEA superintendent or another district administrator overseeing the OTES pilot, and principals and teachers who were participating in the OTES pilot. While visiting the LEAs, evaluators also conducted individual interviews with non-participating principals at 42 percent of the sites and non-participating teachers at 58 percent of the sites. Interviews with non-participating principals and teachers were conducted in lieu of focus groups because of the highly sensitive nature of the topic under investigation. Even with individual interviews, people raised concerns about confidentiality. The evaluators assured each interviewee that they and their documentation would not be identifiable in any report. The interview approach allowed individuals to speak more freely than in a focus group setting.

EXHIBIT 3  
TYPES OF LEA INFORMANTS INTERVIEWED

Types of LEA Interviewees	% of LEAs Visited
District Administrators	92%
OTES Principals	92%
OTES Teachers	92%
Non-OTES Principals	42%
Non-OTES Teachers	58%

In addition to interviews with LEAs, the evaluators conducted a review of OTES implementation documentation to examine types of OTES evidence in participating districts. LEAs were asked to show the evaluator a copy of their comprehensive communication plan for the evaluation system, at least one teacher evaluation written for two pilot teachers and two evaluations written by the same evaluator during the previous year, LEA policies and/or procedures related to the teacher evaluation process

including the use of student achievement data as a part of the teacher evaluation process, and contract language related to the use of teacher evaluations for compensation, placement, or retention decisions, if available. If the LEA district developed its own evaluation system, the evaluators requested a copy of the documents describing the model, including any rubrics or evaluation descriptions.

This case study report focuses on the results of site visits with regard to benefits and challenges that teachers, principals, and other OTES pilot team members are experiencing while implementing the new teacher evaluation system. The report also describes adaptations LEAs are making or considering for the Ohio statewide model to meet the needs of their local school community.

## 4.2 Case Study Findings

### RESEARCH QUESTION 1: IMPLEMENTATION

#### 1a. To what extent are teachers, administrators and union leaders involved in the design and implementation?

The follow up survey will examine the extent to which teachers, administrators and union leaders are involved in the design and implementation of the OTES at LEA pilot sites in Ohio.

#### 1b. What is the fidelity in relation to the project plan?

LEAs that were visited by the evaluators are actively involved in implementing OTES components to the extent shown in *Exhibit 4*. LEAs are giving most attention to implementing the teacher evaluation components, identifying how well teachers are meeting standards 1 - 5 through formal observations, post-observation progress conferences, written observation reports and teachers' use of self-assessment tools and setting SMART goals. The least attention has been given to evaluating teachers' collaboration, communication and professionalism for meeting standards 6-7 and documenting informal walkthroughs.

EXHIBIT 4  
TEACHER EVALUATION COMPONENTS IMPLEMENTED

Teacher Evaluation Components Implemented	N = 12	% of LEAs Visited
Standard 1: Understands student learning and development		100%
Standard 2: Understands content areas taught		100%
Standard 3: Uses student assessments to inform instruction		100%
Standard 4: Plans and delivers effective instruction		100%
Standard 5: Creates environment that promotes learning		100%
Standard 6: Collaborates/communications with community		25%
Standard 7: Takes responsibility for professional growth		25%
Initial goal-setting conference		92%
Teacher performance formal observation		100%
Teacher professional project		0%
Post-observation progress conference		92%
Teacher performance informal walkthroughs		33%
Cumulative observation rating		42%
Evaluator's written report for teaching standards 1-5		75%
Evaluator's written analysis of teacher's collaboration, communication and professionalism standards 6-7		17%
Teacher's written lesson reflection		33%
Teacher's professional growth plan and PD resource use		17%
Teacher's assessment of student growth and achievement		42%
OTES Improvement Plan for marginal performance		0%
Year-end summative conference		8%
Using the standards for the teaching profession for self-assessment tool		100%
Ohio Continuum of Teacher Development Resource Tool		0%
Data collection tool for communication and professionalism		8%
Teacher self-assessment summary tool		67%

### 1c. To what extent were comprehensive communication plans developed and successfully utilized?

Exhibit 5 shows the percent of case study sites that had the various types of documentation available for review. Written teacher evaluations were the most prevalent type of documentation with 75 percent of LEAs able to show documents as evidence of their OTES pilot efforts. At 67 percent of the case study sites, LEAs had implemented policies and procedures related to the new teacher evaluation process. At 75 percent of the LEA sites where documentation was not available, administrators explained that a comprehensive communication plan would be created only after the OTES guidelines were fully developed at the state level and adopted by their local school boards.

EXHIBIT 5  
TYPES OF OTES DOCUMENTATION

LEA OTES Implementation Documentation	% of LEAs Visited
Policies and procedures related to the teacher evaluation process	67%
Policies and procedures related to the use of student achievement data as a part of the teacher process	42%
Written teacher evaluation for each teacher participating in the pilot during the 2011-12 school year	75%
Two teacher evaluations written by the same evaluator during the 2010-11 school year	50%
Contract language related to the use of teacher evaluations for compensation, placement, or retention decisions	58%
Evaluation model if locally developed or if OTES has been modified by the LEA	33%
Comprehensive communication plan	25%

During interviews, teachers and principals reported spending more time this pilot year communicating during one-on-one meeting conferences about teachers' instructional practices, and strengths and areas for improvement than they had in previous years. Teachers find high value in doing the self-assessment and including it as part of the discussion with their evaluator. Although not required, most teachers reported sharing their self-assessment with their principal. Teachers and principals reported a clearer, shared understanding of the teacher evaluation criteria based on indepth examination of the OTES rubric and individual teacher's rubric scores following formal observations. In this regard, OTES is fostering more meaningful communication about teacher effectiveness between teachers and principals.

### 1d. What were the best practices of the most effective implementers?

The evaluation uses the Ohio Standards for the Teaching Profession as the basis for defining effective teaching and examines the impacts of using OTES options on teachers' professional growth and development, administrative behavior and school/LEA processes. It is premature to identify the best practices of the most effective implementers since none of the case study sites have fully implemented the OTES model and effectiveness data is not yet available.

## RESEARCH QUESTION 2: IMPACT ON TEACHER EFFECTIVENESS AND BEHAVIOR

### 2a. What student achievement and growth measures are LEAs using in the OTES pilot?

Ninety-two percent of the LEAs visited are using EMIS reports, 75 percent are using classroom sociograms, and 42 percent are using student growth measures (e.g. Battelle reports, Terra Nova, and short cycle assessments). During interviews, teachers participating in the pilot and non-participants alike verbalized a general anxiety stemming from confusion about the definition of and appropriate interpretation of VAM and student growth scores. Teachers and some administrators expressed concern about how VAM scores relate to teachers' performance scores and how much weight VAM scores will carry in decisions about teachers' employment and pay scales under the OTES model. They also expressed concern about using reading scores to rate effectiveness of art teachers, for example.

Evaluators found a general lack of understanding among those at LEAs about the state-of-the art of VAM metrics, which are undergoing intense research and development. They expressed concern that the OTES model does not provide clear guidelines for the VAM formula or sources of value-added scores. There is a lack of understanding that student growth scores can include metrics for school and learner characteristics and measures for groups of teachers who collaborate in teaching student cohorts. There also is a lack of understanding that multiple years of data are compiled into VAM scores. The confusion about VAM and student growth at LEAs is not surprising given the current widespread technical debates among policymakers and measurement experts.

There was also both concern and relief that the state plans to allow individual LEAs to define the VAM used in their setting. Those who expressed concern mentioned that it might not be "fair and equitable" across the state and teachers in one area might be being held to a higher or different standard than those in a neighboring district. Those who expressed relief indicated that they would rather be held to locally defined standards than state-wide standards that might not be appropriate for their populations. Teachers and administrators in both groups were unsure how the student growth measures were going to be developed in their LEA and very concerned about having salaries tied to the new evaluations.

### 2b. What were the intended and unintended consequences on instructional practices?

Teachers and principals interviewed during the site visits reported shifts in instructional practice that they attribute to participating in the OTES pilot. The OTES has changed teachers' classroom practices in what they perceive to be positive ways and can be a useful a tool for developing teacher effectiveness. The reported shifts in classroom practice are:

- ◆ More awareness and timely analysis of student data for progress monitoring.
- ◆ More awareness of the need for differentiated instruction, including re-teaching areas of weakness in student learning based on formative assessment data.
- ◆ More attention given to monitoring students' learning progress related to content standards.
- ◆ More active and engaged learning among students, increases in student motivation, and fewer behavioral problems among learners in pilot classrooms.
- ◆ More use of self-directed learning activities and methods to stimulate students' taking responsibility for their own learning as part of differentiated instructional approaches.

### RESEARCH QUESTION 3: IMPACT ON STUDENT ACHIEVEMENT

The case studies investigated the extent to which LEAs have been able to identify impacts of the new teacher evaluation on student achievement and found that it is too earlier in the OTES pilot to draw conclusions about impacts on student achievement or value-added impacts of teachers on student growth. It is also too earlier to make inferences about the impacts of OTES on teacher performance in terms of professional growth although self-reports from teachers, principals and administrators indicate it is having a positive impacts in the areas described above under Research Question 2b.

### RESEARCH QUESTION 4: IMPACT ON ADMINISTRATIVE BEHAVIOR AND SCHOOL/LEA PROCESSES

#### 4a. Have LEA policies and procedures changed as a result of implementing the new teacher evaluation methods?

The evaluators asked to review a copy of documentation of new incentives or policies and procedures at case study sites where available. *Exhibit 6* shows the results of document review conducted by the evaluators during site visits. The majority of the LEAs visited have not yet implemented changes in policy or procedures related to the OTES pilot. The categories used to score the policies and procedures are the following:

- ◆ No Plan – the LEA had not yet started to create any draft policies or procedures relative to the new model of teacher evaluation.
- ◆ Draft Plan – the LEA had started to draft policies or procedures.
- ◆ Partial Plan – the LEA had adopted some policies or procedures, but was not finished.
- ◆ Full Plan – the LEA had adopted revised policies and procedures in support of the new model of teacher evaluation.

As shown in *Exhibit 6*, most districts have not yet started to create policies and procedures in support of the new model of teacher evaluation.

EXHIBIT 6  
LEA POLICY AND PROCEDURES SUPPORTING OTES

LEA Policy and Procedures Supporting OTES	No Plan	Draft Plan	Partial Plan	Full Plan
LEA has policies and procedures that reinforce the teacher evaluation process.	75%	8%	8%	8%
LEA has policies and procedures that support using student achievement data as a part of the teacher evaluation process.	67%	17%	8%	8%
LEA has contract language related to the use of teacher evaluation results in compensation, placement and retention decisions.	83%	0%	0%	17%

#### **4b. To what extent has the new teacher evaluation model impacted professional development?**

Teachers and principals reported finding high value in the SMART goal process which fosters more reflective teaching practices among pilot participants and informs professional development choices more aligned to professional growth needs. Both teachers and administrators reported expecting to revisit the SMART goals in the final evaluation conference, but found that the document did not include any reference to the SMART goals. Most people believed that goal setting should remain as a required element in the OTES model.

Those interviewed also reported that OTES creates a need for teachers' in-depth training on SMART goals, assessment literacy and data-driven practices, differentiated instructional methods, and methods for fostering self-regulated learners. Administrators who would like to use peer coaches to corroborate formal observations see a need for evaluation training to establish inter-rater reliability among peer evaluators and administrator evaluators. In addition, administrators expressed a need for more in-depth training on a regular basis to calibrate their scoring of teacher performance and ensure inter-rater reliability across buildings and LEAs.

#### **4c. How has the teacher evaluation pilot affected the alignment of classroom/building/LEA process and performance outcomes?**

With regard to alignment of OTES with classroom/building/LEA processes, case study findings indicate LEAs face the following challenges or concerns about implementing the current version of OTES district-wide:

- ◆ OTES shifts observations from being unannounced to announced. Many administrators viewed announced events as “canned shows” that any certified teacher can master and produce because the lessons are well planned, but may not be an accurate snapshot of a teacher’s daily classroom practice. They find greater value in unannounced observations and walkthroughs.
- ◆ Walkthroughs are part of OTES, but in some districts, collective bargaining agreements preclude walkthroughs from being used for teacher evaluation purposes.
- ◆ The current OTES model requires coordination of numerous documents that are on paper. LEAs would prefer a more streamlined process in an easy to access electronic format that will allow both teachers and administrators to post, view and rate evidence of teacher effectiveness (e.g. SMART goals, lesson plans, parent letters, student work, observation results, etc.) and hold documents “in-progress” until they are completed. They want a structure that supports both walkthroughs and formal observations, and connects ratings and observation notes more easily to the rubric categories.
- ◆ The amount of time principals need to spend on conducting a full OTES evaluation for a single teacher is viewed as not feasible to scale up for all staff. Administrators view the current OTES as a valuable comprehensive way to evaluate new teachers and those with marginal performance. For proficient teachers, administrators would like modifications to the OTES model including the option to evaluate teachers through progress monitoring for SMART goals, professional development projects, or teacher leadership activities and only conduct the full OTES evaluation on a biennial or other rotation.
- ◆ There is a general lack of understanding among LEAs that, as a pilot site, they have the “green light” to make modifications to the OTES model. Instead, they are highly focused on OTES compliance.

### 4.3 Preliminary Recommendations and Next Steps

The OTES follow up survey will investigate how prevalent the findings from the case studies are throughout Ohio with regard to implementation, impacts on LEA processes and sustainability. It is the opinion of the evaluators that it is too early to adequately describe impacts of OTES on teacher effectiveness and student achievement. The latter is a topic for future evaluation.

#### BEST PRACTICES

##### Literature Review of Trends in Teacher Evaluation Methodology

Across the United States, policymakers, measurement researchers and those in the teaching profession are re-examining and experimenting with innovative ways to evaluate teacher effectiveness. The following review of research and current practices in teacher evaluation methodology was conducted to identify strategies and lessons learned from current teacher evaluation initiatives. Approaches to teacher evaluation that are underdevelopment are summarized with recommendations to inform future refinement of the Ohio Teacher Evaluation System (OTES).

Since 2010, laws and policies governing teacher evaluation have been undergoing rapid changes in the United States. The primary shifts have been from teacher evaluation methods based on a review of teacher behavior and salary and tenure decisions based on seniority and degrees earned to a focus on student learning outcomes. Before 2010, only four states were using student achievement as a predominant influence in how teacher performance was assessed. By 2011, the number had jumped to 13 with 10 additional states including student achievement as a small percent of teacher evaluations, according to the recent report, *“State of the States: Trends and Early Lessons on Teacher Evaluation and Effectiveness Policies”* from the National Council on Teacher Quality (October, 2011).

In September 2011, the federal government directed that states wanting relief from the No Child Left Behind (NCLB) law could apply for a waiver from the law’s tough-to-meet requirements for student achievement in reading and math. To qualify for a waiver, however, states must define how they would use teacher and principal evaluations to make personnel decisions. Initially, 11 states applied for waivers, and an additional 28 states said they planned to seek waivers. In addition to the NCLB waivers, the Race to the Top (RttT) competition motivated states to adopt new evaluation systems in order to win the grant competition. While these policy changes have caused a major shift in local accountability systems for teachers and principals, the use of value-added scores are at the center of the debate about how to best measure teacher effectiveness. Driven by policy mandates, educational systems are beginning to adopt value-added methods (VAM) for evaluating teachers even before researchers have concluded investigations into how best to measure the impact of teaching practices on student growth.

Recently, the Education Commission of the States tracked 18 state legislatures that had modified teacher tenure or continuing contract policies. One of the most notable was Idaho, where legislators enacted a bill banning tenure for new teachers and other certified employees. While some states’ leaders waged intense battles with teachers’ unions, Illinois changed its approach to teacher tenure with less conflict. Governor Pat Quinn signed into law a measure that links educators’ tenure, hiring, and job security to performance, rather than to seniority. The Illinois law makes it easier to remove an educator from the classroom for continuously poor performance. Tennessee and Colorado are among the states

taking an early role in implementing VAM into teacher evaluation systems. These two states have aggressive laws and have implemented practices using VAM for teacher evaluation.

The *Tennessee Comprehensive System* is one of the more established systems using student achievement data in teacher evaluations. In 2011-12, student test-score growth will count for 35 percent of a teacher's year-end evaluation. Districts will use the data to decide which teachers receive tenure and which are let go. An additional 15 percent of a teacher's evaluation is made up of achievement measures chosen by the district, and 50 percent is based on classroom observations and other measures. Under the Tennessee 5-point rating system, teachers defined as a 3, or "at expectations," are those whose students make at least one grade level of gain on the state's test. To receive tenured status, a new teacher's students must demonstrate more than one year's gain for two years. William Sanders, a former University of Tennessee researcher who now works for SAS, a private business-intelligence company, developed Tennessee's value-added formula which does not factor in any individual student characteristics. Previously, teachers in Tennessee were evaluated only once every five years. Under the new system, principals are required to spend from 60 to 90 minutes in a teacher's classroom annually, with the amount of time in classrooms dependent on a teacher's experience. For veteran teachers, principals must conduct four 15-minute observations over the course of the school year [see <http://www.tn.gov/firsttothetop/programs-committee.html>].

The *Colorado State Council for Educator Effectiveness* (CSCEE) was established in 2010 as part of Senate Bill 191 to design a value-added model to provide 50 percent of the state's teacher evaluation data. The Colorado law requires LEAs to conduct performance evaluations for all teachers and principals at least once each school year. At least half of each teacher's and principal's evaluation must be based on multiple measures of students' academic growth, including the state test. Teachers' performance effectiveness ratings must be used before seniority when considering district-level layoffs. Colorado law also requires LEAs to consider student mobility and the numbers of students with disabilities or at risk of failing school in its VAM formula.

Colorado implemented a pilot program for new teacher evaluation methods during the 2011-2012 school year, the results of which are not yet available. However, in an April 2011 report, the CSCEE said it had established a Technical Advisory Group of local and state stakeholders to assist with development of recommendations and is conducting an evaluation with LEAs that are piloting new teacher evaluation methods and research review of best practices. The April 2011 recommendations from the CSCEE emphasized alignment of LEA methods with the Colorado Teacher Quality Standards and called for common, statewide technical guidelines for selection and use of valid and reliable measurements including how to translate assessment data into growth scores for evaluation purposes. The CSCEE also called for protecting educator evaluations as private data within the state accountability system.

The CSCEE recommends that the state develop summative assessments for 70 percent of the teachers who currently teach untested subjects or grades and align state policies to quality standards for licensure, accreditation of preparation programs, approval of induction programs, professional development, and educator recognition criteria. The CSCEE recommends that the state provide one complete educator evaluation model with supporting measurement tools and exemplars of LEA evaluation practices. They also recommend that LEAs be encouraged to attribute student growth to teams of educators instead of to individual teachers and to use student and parent/guardian perceptual data as part of teacher evaluation. The CSCEE reported findings from a study of LEA costs to implement the new teacher evaluation system. The estimated annual cost was \$531 for effective teachers and

\$3,783 for ineffective teachers who required more supervision and remediation. The study estimated an additional one-time cost of \$53 per student in the first year of implementation.

In addition to Colorado and Tennessee, the *Delaware Performance Appraisal System* is a statewide educator evaluation system that has been in place since 2008 but is undergoing significant modifications which are being piloted during the 2011-12 school year. The teacher evaluation was initially based on Charlotte Danielson's *Enhancing Professional Practice: A Framework for Teaching (2nd Edition. 2007)*.

The recent changes to teacher evaluation are summarized in a policy briefing from the Delaware Department of Education. (See <http://www.doe.k12.de.us/rttt/files/initiatives/DPASII.pdf> ) In the Delaware model, a school-wide assessment measure, a student cohort assessment measure, and a teacher-specific assessment measure will be combined on a 100-point scale for the student growth component of the teacher effectiveness rating.

The school-wide assessment measure will be used for all teachers and specialists and accounts for 30 points. Each educator will receive either the Delaware Comprehensive Assessment System (DCAS) reading score across all grades in school or the DCAS mathematics score across all grades in school, with the determination based on which one shows the most positive result using DCAS Adequate Yearly Progress (AYP) scores.

For the student cohort assessment measure, districts in Delaware are asked to use the fall-spring growth based on students' instructional scores on the DCAS, accounting for 20 points. A student cohort could be specific to a grade level, subject area or a student-based cohort within a test grade/subject area. For example, a counselor may identify a subset of students with frequent absences and focus on math with that group of students. The teacher-specific assessment measure, which counts for 50 points, is a non-DCAS measures tied directly to the teacher or specialist's current teaching assignment and to be approved by the Secretary of Education.

In Georgia, 26 local education systems are participating in the state's Race to the Top grant to pilot its new state-wide teacher evaluation system, called Teacher Keys, beginning in January 2012. The system includes student performance data, student surveys about teachers and teacher surveys about principals, and administrators' two classroom observations. Pilot sites in Georgia will field test the new electronic platform that provides web-based access to the evaluation process guides, templates, and support materials.

Beginning in 2012-13, participating sites will have access to a data warehouse for all observation records, documentation to supplement and support those observations, student survey and growth data, and other relevant information. An electronic record will be maintained of all components of the evaluation system, including orientation, familiarization, self-assessment, TAPS formative and summative documents, student surveys, Student Learning Objective (SLO) data and evaluation, student growth percentile data and calculations. During the pilot, functionality of the platform will be limited with linkages to student data expected in the next few years. For more details, see the Georgia State School Superintendent's explanation of the Teacher Keys Evaluation System Pilot online at [http://www.rabun.k12.ga.us/RTTT\\_Rabun.pdf](http://www.rabun.k12.ga.us/RTTT_Rabun.pdf)).

## Perspectives on VAM from Measurement and Research Experts

In a joint statement from the National Council for Measurement in Education, the American Psychology Association, and the American Education Research Association researchers explain,

Tests valid for one use may be invalid for another. Each separate use of a high-stakes test, for individual certification, for school evaluation, for curricular improvement, for increasing student motivation, or for other uses requires a separate evaluation of the strengths and limitations of both the testing program and the test itself.

The National Research Council's Board on Testing and Assessment (2009) made similar claims and warned against using standardized test scores to evaluate the effectiveness of teachers without first determining the degree of alignment between the test, content standards, instruction and curriculum. The Council's Board on Testing and Assessment (BOTA) asserts,

BOTA has significant concerns that the [federal] Department's proposal places too much emphasis on measures of growth in student achievement (1) that have not yet been adequately studied for the purposes of evaluating teachers and principals and (2) that face substantial practical barriers to being successfully deployed in an operational personnel system that is fair, reliable, and valid. . . . The term "value-added model" (VAM) has been applied to a range of approaches, varying in their data requirements and statistical complexity. Although the idea has intuitive appeal, a great deal is unknown about the potential and the limitations of alternative statistical models for evaluating teachers' value-added contributions to student learning. BOTA agrees with other experts who have urged the need for caution and for further research prior to any large-scale, high-stakes reliance on these approaches (e.g., Braun, 2005; McCaffrey and Lockwood, 2008; McCaffrey et al., 2003).

A primary concern among researchers is that it is technically inaccurate to assign causality of student learning outcomes to teachers in real world classrooms as teachers and students are not randomly assigned to class groups. Technical issues that complicate the meaning of value-added scores include this lack of random assignment of teachers to schools and students to teachers. Current VAM techniques do not control for those differences and therefore VAM scores are not comparable between teachers who work with different populations. In addition, value-added scores may be affected by student motivation and parental support (BOTA, 2009). Linn (2008) concludes: "As with any effort to isolate causal effects from observational data when random assignment is not feasible, there are reasons to question the ability of value-added methods to achieve the goal of determining the value added by a particular teacher, school, or educational program."

Darling-Hammond (et al., 2012) further explain: "Value-added models enable researchers to use statistical methods to measure changes in student scores over time while considering student characteristics and other factors often found to influence achievement. In large-scale studies, these methods have proved valuable for looking at factors affecting achievement and measuring the effects of programs or interventions." Many of the leading researchers in the area of teacher effectiveness and educational measurement concur (Baker et al., 2010; Braun, 2005, 2011; Newton, 2011) that with respect to using value-added measures of student achievement for evaluating individual teachers, there is strong research evidence that suggests that high-stakes, individual-level decisions, or comparisons across highly dissimilar schools or student populations, should be avoided.

The term “student growth” is defined differently for different value-added models. As Darling-Hammond (et al., 2012) explain, research reveals that a student’s achievement and measured gains are influenced by much more than by the work of any individual teacher. Others factors include:

- ◆ School factors such as class sizes, curriculum materials, instructional time, availability of specialists and tutors, and resources for learning (books, computers, science labs, and more)
- ◆ Home and community supports or challenges
- ◆ Individual student needs and abilities, health, and attendance
- ◆ Peer culture and achievement;
- ◆ Prior teachers and schooling, as well as other current teachers
- ◆ Differential summer learning loss, which especially affects low-income children
- ◆ The specific tests used, which emphasize some kinds of learning and not others, and which rarely measure achievement that is well above or below grade level.

Most of these factors are not actually measured in current value-added models, and the teacher’s effort and skill, while important, constitute a relatively small part of this complex equation. As a consequence, researchers have issued the following cautions about adopting VAM models to accurately measure teacher effectiveness:

- ◆ Current value-added models of teacher effectiveness are highly unstable as research shows a teacher’s effectiveness rating can differ substantially from class-to-class or year-to-year and across different VAM statistical models.
- ◆ Current value-added models do not account for disproportionate numbers of students with poor attendance, unstable home environments, low parental involvement, lack of internal motivation, and foreign language background, which cause misestimates of teachers’ effectiveness and disincentives for teaching students with the greatest needs.
- ◆ The accuracy of value-added models is greatly improved with the random assign of students to teachers; however, the likelihood of random assignment is low in our educational systems; and
- ◆ Current value-added ratings do not account for the complexity of influences such as multiple teachers, school conditions, prior teachers, and classroom groupings.
- ◆ Often the VAM ratings generated by today’s rudimentary methods do not correlate with teacher evaluation ratings assigned by skilled observers. In terms of statistical validity, teacher evaluations need multiple measures that all point to the same conclusion.

Harris (2011) explains that “there are considerable errors in value-added measures” which need to be addressed and caution is needed in how VAM scores are used within teacher performance systems. In addition, VAM scores are based on the bell curve which inherently ranks scores so that some teachers will always be below average and some above average even if the difference between them is insignificant. Harris also explains that any accountability system should focus on holding teachers accountable for what they control, which is goals, lesson plans, progress monitoring of students, differentiated instruction and classroom management. Students also have control in the classroom with regard to their behavior, motivation, receptiveness to learning, peer influences, attendance, health and nutrition that influence teaching. School communities control factors such as class size, class grouping, administrative leadership and support, curriculum resources, support staff, and parental involvement

that influence teaching. Larger societal trends, current events, and funding all impinge on student behaviors and teaching practices in the classroom.

Prior to 2010, the most common methods for teacher evaluation have been classroom observations and evaluations by administrators, teacher portfolios documenting a teaching behaviors and responsibilities, and peer review (Hinchey, 2010). Darling-Hammond (et al., 2012) also point out that countries like Singapore include a major emphasis on teacher collaboration in their evaluation systems. This kind of measure is supported by studies which have found that stronger value-added gains for students are likely in schools where teachers work together as teams and have higher levels of teacher collaboration. Just as No Child Left Behind sought to address problems with low expectations for student performance and led to focus on narrow curriculum and teaching to the test, Hinchey warns that the VAM movement's focus on test scores also will lead to undesirable consequences that undermine the goal of developing excellence in the teacher workforce. Harris (2011) reviewed the many models of value-added measures under development by assessment and measurement experts and concluded that VAM is not suitable for use in evaluating individual teachers, but when applied appropriately, school value-added measures are useful for accountability.

### **Merit Pay and Bonus Incentives**

Another area of debate is the merit pay incentives that states and districts are packaging with the new teacher evaluation systems. School districts in at least 42 states have some form of merit pay, according to the National Center on Performance Incentives at Vanderbilt University. In a study of the District Awards for Teacher Excellence (DATE) merit pay program in Texas, Springer and Lewis (2010) concluded that there was a correlation between schools' voluntary participation in the merit pay program, increased teacher retention, and improvements in students' test scores in some, but not all, participating schools.

In 2011, Indiana passed a new law mandating that test performance of students be factored into teacher pay raises. This move represents a major shift away from the pay scales that award pay raises based primarily on a teacher's years of experience and the academic degrees they earned. At Indiana's Wayne Township Schools, administrators are proposing a new compensation system based on a seven-point award system (Elliott, S. & Butrymowicz, S., 2012). Under the plan, teachers receive one point each for years of teaching experience, degrees attained, professional leadership, attendance, and up to three points for performance based on a formal evaluation. Within the proposed evaluation method, student test scores carry 20 percent weight. Eighty percent of a teacher's evaluation is based on observations of their work. The school district's superintendent wanted the compensation system to motivate all teachers to meet their students' academic needs, not just those in a particular subject area or hard-to-fill position.

Berry and Eckert (2012) point out that incentive programs that merely pay teachers for student test scores produce limited results; other incentives can produce better outcomes and can be used to spread expertise to colleagues. In their review of the empirical evidence, Berry and Eckert note that "teacher incentive proposals are rarely grounded on what high-quality research indicates are the kinds of teacher incentives that lead to school excellence and equity." For example, the authors note that "empirical evidence, including large-scale studies and an increasing number of teacher testimonies, suggest that working conditions are far more important than bonuses." Moreover, those important working conditions go well beyond the issues of time, class size, and the length of the workday. The working

conditions that appear to support improved teacher and student performance are those that allow teachers to teach effectively, including:

- 1) Principals who cultivate and embrace teacher leadership.
- 2) Time and tools for teachers to learn from each other.
- 3) Specialized preparation and resources for the highest needs schools, subjects, and students.
- 4) The elimination of out-of-field teaching assignments.
- 5) Teaching loads that are differentiated based on the diversity and mobility of students taught.
- 6) Opportunities to take risks.
- 7) Integration of academic, social, and health support services for students.
- 8) Safe and well-maintained school buildings.

In addition, missing from virtually all of the currently in-vogue strategies to give teachers incentives to improve achievement is an understanding of how incentives could be used to reward teachers who spread their expertise to their colleagues. Teachers have long been organizationally “siloeed” from each other. Berry and Eckert (2012) point out that strategic compensation could be used to reward teachers who collaborate, not compete, with their colleagues in helping them teach more effectively.

Still, many educators criticize performance pay plans, arguing that the promise of more money will not make teachers work harder. Performance appraisal systems that focus heavily on test scores can have negative consequences, such as cheating and narrowing of learning opportunities when teachers focus on test content (Chetty, Friedman, Rockoff, 2011). Jabbar (2011) argues for an approach that would incorporate psychological knowledge about human behavior related to intrinsic motivation and decision-making to enhance student performance measures in education rather than external motivation incentives, such as merit pay.

Dobbie & Fryer (2011) did an intensive mixed methods study of successful charter schools in New York and found five strategies in effective schools: give frequent feedback to teachers, use loads of data on individual students to guide their instruction, employ heavy tutoring, increase instructional time, and maintain very high expectations. These strategies are school-wide and require administrative support to implement in the classroom and teachers need formative evaluation feedback during the school year, access to high quality data systems and short cycle assessments aligned to high quality benchmarks and outcome standards for content area learning, time to analyze student learning data on a regular basis, and time to plan and implement differentiated instruction which may require the support of additional teachers or resource professionals for students performing at the low-end and high-end margins. These types of strategies require a coordinated, school-wide effort. Incentives and methods that pit individual teachers against each other or that do not support professional collaboration and communications among teaching staff and administrators can be counterproductive to fostering students’ academic growth. Darling-Hammond (et al., 2012) concur that a viable alternative to use of VAM in teacher evaluation is to focus on teachers’ planning processes and use of evidence-based instructional strategies proven to be effective for fostering students’ academic growth.

## 4.4 Conclusions and Recommendations from the Case Study and Review of Best Practice Literature

The use of value-added measures to evaluate teachers raises many technical and ethical issues that will be debated for years to come as assessment experts and educators investigate correlations and causal links between teacher behaviors and student growth. In the meantime, principals, school districts and teachers struggle to meet new mandates set up by new laws that call for a student growth score to constitute up to 50 percent of a teachers' evaluation rating. However, there are indications that policymakers are pausing to rethink the use of VAM within teacher evaluation systems. For example, the Elementary and Secondary Education Act (ESEA) of 2011 scaled back its original mandate for student growth measures in teacher evaluation systems at the federal level.

Current research indicates that student achievement and measured gains are influenced by much more than an individual teacher. Other school factors influencing student achievement include class sizes, instructional time, availability of specialists and tutors, degree of alignment of curriculum resources with content standards and effective assessment measures, previous and other teachers, school leadership. Out-of-school factors influencing student achievement include level of supports or challenges in the home environment, individual learner needs and abilities and intrinsic motivation, personal and family health, and school attendance rate, peer culture, and differential summer learning loss.

VAM researchers agree that the lack of random pairing of students and teachers makes causal attributions a fundamental flaw of current VAM formulas. There is a wide-spread interest in resolving this technical issue. However, until more robust value-added formulas become available, experts strongly caution against using VAM data to make high-stakes decisions about critical issues such as teacher pay, tenure, awarding of teaching licenses and ranking of teacher effectiveness.

A major finding of the review of research and practices related to teacher evaluation methodology is the understanding that while educational policy is well-intended in seeking to link teachers' classroom practices to student learning, there is much work to be done to resolve technical issues with current value-added methodologies for doing so. Few dispute that teachers have a significant impact on student learning. The debate is how to best measure that impact and on that point there is yet to be consensus among policy makers and researchers. It is recommended that educators conduct pilot studies of VAM and take a more evidence-based approach to including student growth data in the evaluation of teachers.

In summary:

- 1) Experts caution against using VAM scores as the basis for high-stakes decisions about tenure, hiring and pay incentives. Experts also caution against giving substantial weight to VAM scores to avoid consequences such as cheating and teaching to the test.
- 2) Educators need to understand how to interpret VAM scores and evidence-based VAM methodologies in order to be properly informed when selecting vendors and weighting VAM scores within the OTES model.
- 3) Longitudinal studies needs to be conducted to verify how VAM scores provided by vendor assessment systems correlate with other measures of teacher effectiveness such as observations, classroom assessments, graduation rates, and college readiness.

- 4) Current value-added methods offer limited value for teacher appraisals and need to be only one of multiple measures used for evaluating teacher effectiveness. Other measures could include SMART goals and progress monitoring, observations and walkthroughs, lesson plans, short cycle assessment results, benchmark assessments, student work and professional collaboration, context measures for special needs, class grouping, school, home and community characteristics.
- 5) Educators should focus on designing an evaluation system that provide useful feedback to individual teachers, teacher cohorts, and school administrators to use teacher effectiveness data to improve instruction.

These findings from the literature review were compared to findings from the OTES case study sites and the following recommendations are made relative to the improvement and scalability of a sustainable OTES model. Amidst the thriving debate about how to evaluate teachers, leaders in policy, education and measurement all agree on one thing: teacher effectiveness can and should be evaluated. The challenge is to make operational the well-intended movement toward a better educational system for Ohio's children.

With regard to using value-added measures in evaluations of teacher effectiveness, the literature reviewed indicates that such measures have significant errors that need to be minimized through statistical research in more depth. It will take time to develop value-added measurements that are valid and reliable enough to use in schools. In the meantime, it would be judicious to:

- 1) Avoid using VAM data for high-stakes decisions about individual teacher's effectiveness.
- 2) Include VAM as only a small percent of teacher groups' evaluation results.
- 3) Conduct rigorous pilot studies of how the results of various value-added statistical models correlate with other measures for student growth and teacher effectiveness.

In addition, it is recommended that OTES not lose sight of the fact that this new teacher evaluation movement in the United States reaches far beyond value-added scores to coalesce a decade or more of work to improve teacher effectiveness through high-quality professional teaching standards, common core content standards, meaningful curriculum and multiple assessments aligned to standards and the accessibility of timely data systems. While the expected outcome is improvement in student academic growth, the state and its districts need to focus on alignment of their standards, policies, resources and assessment measures while they improve the teacher evaluation system. Districts need to provide teachers and educational leaders with professional development to appropriately and effectively implement supports for improving educational opportunities for students.

Educators interviewed at case study sites generally do not understand the complexity of value-added measures, how a VAM score is derived, how to interpret it in terms of improving their teaching practice, or how to interpret VAM for parents. These are important topics that need to be address through action research, professional development, and communication plans before student growth scores are introduced to the public.

In the end, teacher evaluation is best grounded in criteria for which teachers have control, such as setting SMART goals, regularly measuring and monitoring the learning progress of students, adjusting classroom practice to address gaps in learning or to accelerate learning through differentiated

instruction, engaging in professional development aligned to needs assessment results, and collaborating with support staff, other teachers and administrative leaders to address the needs of all students within a school community.

## Best Practice Bibliography

Baker, E., Barton, P., Darling-Hammond, L., Haertel, E., Ladd, H., Linn, R., Ravitch, D., Rothstein, R., Shavelson, R., Shepard, L. (2010) *Problems with the use of student test scores to evaluate teachers*. Washington, D.C.: Economy Policy Institute. [Available online at <http://www.epi.org/publications/entry/bp278>].

Barlevy, G. & Neal, D. (2012). Pay for percentile. *American Economic Review* [forthcoming].

Berry, B. & Eckert, J. (2012) *Creating Teacher Incentives for School Excellence and Equity*. Bolder, CO: National Education Policy Center. [Available online at <http://nepc.colorado.edu/files/NEPC-PB-TchrPay.pdf>].

Braun, H. (2005). *Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models*, Educational Testing Service Policy Perspective, Princeton, NJ [Available online at <http://www.ets.org/Media/Research/pdf/PICVAM.pdf> ].

Board on Testing and Assessment. (2009). Letter Report to the U.S. Department of Education on the Race to the Top Fund. Washington, D.C.: National Research Council. {Available online at [http://www.nap.edu/catalog.php?record\\_id=12780](http://www.nap.edu/catalog.php?record_id=12780)}.

Chetty, R., Friedman, J.N., Rockoff, J.E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. Cambridge, MA: National Bureau of Economic Research.

Danielson, C. (2008). *The Handbook for Enhancing Professional Practice: Using the Framework for teaching in your school*. Alexandria, VA: Association for Supervision and Curriculum Development (ASCD).

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, 93 (6): 8-15.

Dobbie, W., & Roland G. Fryer, J. (2011). Getting Beneath the Veil of Effective Schools: Evidence from New York City. NBER Working Paper No. 17632. Cambridge, MA: National Bureau of Economic Research.

Dobbie, W., & Roland G. Fryer, J. (2011). Are High-Quality Schools Enough to Increase Achievement Among the Poor? Evidence from the Harlem Children’s Zone. *Public Policy*, 3(November), 158–187.

Elliott, S. & Butrymowicz, S. (April 9, 2012). Questions abound as districts shift to merit pay for teachers. *The Hechinger Report* [available online at [http://hechingerreport.org/content/questions-abound-as-districts-shift-to-merit-pay-for-teachers\\_8260/](http://hechingerreport.org/content/questions-abound-as-districts-shift-to-merit-pay-for-teachers_8260/)].

Harris D.N. (2011). *Value-added measures in education*. Cambridge, MA: Harvard Education Press.

Hinchey, P.H. (2010). *Getting teacher assessment right*. Boulder, CO: National Education Policy Center [Available online at [http://nepc.colorado.edu/files/PB-TEval-Hinchey\\_0.pdf](http://nepc.colorado.edu/files/PB-TEval-Hinchey_0.pdf)]

[Jabbar](#), H. (2011). The behavioral economics of education: New directions for research. *Educational Researcher*, 446-453.

Linn, R.L. (2008). *Measurement issues associated with value-added models*. Paper prepared for the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC, November 13-14. [Available online at [http://www7.nationalacademies.org/bota/1VAM\\_Workshop\\_Agenda.html](http://www7.nationalacademies.org/bota/1VAM_Workshop_Agenda.html) [September2009].

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010) Value-Added Modeling of Teacher Effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives*, 18 (23). [Available online at <http://epaa.asu.edu/ojs/article/view/810>].

Marzano, R.J. (2011). *The Marzano teacher evaluation model*. Englewood, CO: Marzano Research Laboratory.

National Academies, Board on Testing and Assessment and the National Academy of Education Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Accountability. (November 13-14, 2008). *Collection of Conference Workshop White Papers*.

Washington, DC: National Academies.

Newton, S.P. (2011). *Predictive Validity of the Performance Assessment for California Teachers*. Stanford, CA: Stanford Center for Opportunity Policy in Education, 2010.

Springer, M.G. & Lewis, J.L. (2010). District Awards for Teacher Excellence (D.A.T.E.) Program: Final Evaluation Report. Policy Evaluation Report. Nashville, TN: National Center on Performance Incentives at Vanderbilt University. [Available online at [http://www.performanceincentives.org/data/files/news/BooksNews/FINAL\\_DATE\\_REPORT\\_FOR\\_NCPI\\_SITE.pdf](http://www.performanceincentives.org/data/files/news/BooksNews/FINAL_DATE_REPORT_FOR_NCPI_SITE.pdf)].