MATHEMATICA Policy Research

REPORT

FINAL REPORT

Measuring Teacher and School Value Added in Oklahoma, 2012–2013 School Year

May 21, 2014

Elias Walsh
Albert Y. Liu
Dallas Dotter

Submitted to:
Office of Educator Effectiveness
Oklahoma State Department of Education
2500 N. Lincoln Boulevard
Oklahoma City, Oklahoma 73105
ATTN: Dr. Kerri White

Submitted by:
1100 1st Street, NE
12th Floor
Washington, DC 20002
Project Director: Elias Walsh
Reference Number: 40249.720

This page has been left blank for double-sided copying.

CONTENTS

This page has been left blank for double-sided copying.

TABLES

This page has been left blank for double-sided copying.

ACKNOWLEDGMENTS

We are grateful to the many people who contributed to this report. First, we would like to thank the Oklahoma State Department of Education (OSDE) for funding the work. Several people at OSDE played key roles facilitating the development of the value-added model described in this report, including Kerri White, Jenyfer Glisson, Ginger DiFalco, and Susan Pinson. Jackie Skapik at Urban Policy Development also made valuable contributions. We also thank the members of the Teacher and Leader Effectiveness Commission, chaired by State Superintendent Janet Barresi, for their guidance as the value-added model was being developed. Most of all, we would like to thank the TLE working groups consisting of over 300 teachers, administrators, and education experts who dedicated many hours to build a comprehensive TLE model that meets the needs of the state of Oklahoma. The members of the Technical Advisory Board organized by OSDE—Andy Baxter, Southern Regional Education Board; Terri Grissom, Oklahoma GEAR UP; Todd Hellman, Battelle for Kids; Betina Jones-Parra, Noble High School; Daniel McCaffrey, Educational Testing Service-Carnegie Foundation; Mark Snead, RegionTrack, Inc.; and Christina Theokas, The Education Trust—also provided valuable suggestions.

At Mathematica Policy Research, Chris Rodger oversaw a team of programmers, including Clare Wolfendale, Swaati Bangalore, Lisa McCusker, Eric Lundquist, Matthew Jacobus, and Alena Davidoff-Gore, who processed the data and provided expert programming. Duncan Chaplin provided valuable comments. John Kennedy edited the report, and Sheena Flowers and Jackie McGee provided word processing and production support.

This page has been left blank for double-sided copying.

I. OVERVIEW

In this report, we describe the value-added model used as part of the state of Oklahoma's Teacher and Leader Evaluation System (TLE). We estimated measures of teacher and school effectiveness based on instruction provided during the 2012–2013 school year. These measures will be provided to teachers and administrators in the spring and summer of 2014.

The 2013–2014 school year is the first of two pilot years for the quantitative components of the TLE system, during which no stakes will be attached to the value-added results. The full implementation of the TLE system will begin in the 2015–2016 school year, incorporating value-added results based on instruction provided during the 2014–2015 school year. At that time, the Oklahoma State Department of Education (OSDE) will combine the value-added results with additional TLE components to produce composite TLE ratings for teachers and administrators. OSDE plans for the value-added results to account for 35 percent of eligible teachers' and administrators' TLE ratings. OSDE intends for educators to use the TLE ratings to promote continuous improvement of instruction and student achievement.

In designing the value-added model, we worked closely with the TLE Commission, a group composed of key stakeholders appointed by the Oklahoma Governor, President Pro Tempore of the Senate, and Speaker of the House of Representatives. The TLE Commission's role is to issue recommendations to the Oklahoma state board of education about the design of the TLE system, including the value-added model. We identified key decisions regarding the value-added model and then made recommendations for each decision. In doing so, we assessed whether various options advanced certain objectives, such as increasing the accuracy of the value-added results. Throughout the process, we coordinated with key staff at OSDE to understand the state's policy context and priorities. In addition to our recommendations, the TLE Commission also sought advice from educator work groups and a seven-member Technical Advisory Board. The state board then made the final decisions about the design of the value-added model based on the TLE Commission's recommendations.

In the rest of this chapter, we provide an overview of value-added methods in nontechnical terms, provide an overview of the value-added model, and discuss the limitations of value-added models. In Chapter II, we present the key decisions about the value-added model approved by the state board. In Chapter III, we discuss the data used in the value-added model, and in Chapter IV, we provide a detailed discussion of how we estimated the teacher value-added model. Finally, in Chapter V, we describe how we translated value-added results to the scale used in the TLE system and how we calculated value-added results for schools and student subgroups.

A. Using value added to measure teacher effectiveness

Value added is a measure of teachers' contributions to students' academic growth. The measure compares the achievement of a teacher's students to an estimate of how the same students would have achieved with an average teacher. The measure is known as value added because it isolates a teacher's contribution from factors that are not in the teacher's control.

The basic approach of value-added models is to compare two test score averages for each teacher: (1) the average actual scores that the students obtained with the teacher and (2) the

average estimated scores that the same students would have obtained with an average teacher. The difference in these two average scores—how the students actually performed with a teacher versus how they would have performed with the average Oklahoma teacher—represents a teacher's value added to student achievement.

The estimated scores that the students would have obtained with an average teacher—sometimes called predicted or typical scores—are called typical-peer scores in the TLE system. OSDE chose this term because it highlights that the scores are estimated by looking at the achievement of students' most similar "peers" in the state—those with similar previous scores on multiple assessments and other background characteristics. Rather than comparing a student's achievement only to a relatively small number of students with identical background characteristics, we used a statistical technique called multiple regression, which simultaneously estimates a relationship between each included background characteristic and achievement. For each characteristic, this technique compares the achievement of students with the characteristic to the achievement of all other students in the state. Because a student's typical-peer score is based on these statewide relationships between background characteristics and achievement, it represents how the student would be predicted to perform with an average Oklahoma teacher.

By comparing actual and typical-peer scores, value-added models enable any teacher to be identified as a high performer, regardless of the baseline achievement levels or background characteristics of the teacher's students. For example, suppose that a grade 6 math teacher has a class of students who, given their background characteristics such as poverty status, disability status, and test scores on the grade 5 math, reading, and science tests (or pre-tests), typically end the year with a score of 750 on the grade 6 math test (or post-test). The value-added model calculates a relative measure of the teacher's effectiveness by comparing this class average typical-peer score to the class average actual post-test score. In this example, if the average actual score is also 750, the value-added model will identify the teacher as an average performer because the typical-peer and actual scores are equal. If the post-test average exceeds this standard, the teacher will be identified as above average; conversely, if the average is lower than the standard, the teacher will be considered below average.

A number of prominent researchers have used value-added models. For example, they have used value-added models to show that teacher effectiveness is associated with the outcomes of students later in life (Chetty et al. 2011). Value-added results have also been shown to closely align with student outcomes in a randomized controlled trial experiment (Kane et al. 2013). Many school districts and states now use value-added models to measure the performance of schools and/or teachers. For example, Mathematica Policy Research has developed value-added models for Pittsburgh Public Schools (Johnson et al. 2012), Pennsylvania State Department of Education (Walsh and Lipscomb 2013), and in Washington, D.C. (Isenberg and Walsh 2013).

B. The value-added model for Oklahoma

Although conceptually straightforward, the production of value-added results that accurately and fairly measure teachers' performance requires (1) the assembly of an analysis file of data from multiple sources and (2) the design of a value-added model that addresses Oklahoma's educational context. We briefly describe the key elements of the analysis file (described more

fully in Chapter III) and then provide an overview of the steps used to estimate value added (see Chapter IV for details).

We developed approaches to estimating value added based on two types of test scores from the 2012–2013 school year: (1) Oklahoma Core Curriculum Tests (OCCTs) in grades 4 through 8 in math and reading; and (2) End of Instruction (EOI) assessments for students in grades 8 and 9 for algebra I, grades 9 through 11 for geometry, grades 9 through 12 for algebra II, and grade 11 for English III. We refer to test scores from the 2012–2013 school year as post-test scores. The value-added model also uses selected test scores from the 2011–2012 school year, which we refer to as pre-test scores. The value-added model yields a value-added result for each subject taught by a given teacher.

Students were eligible to be included in the model if they had a post-test score and a pre-test score from the previous grade in the same content area. For example, for students with grade 5 math post-test scores, the analysis file includes only those students who also have grade 4 math pre-test scores. For a student with a grade 10 geometry post-test score, the analysis file includes only students with a grade 9 pre-test for a subject in the math content area, such as algebra I. We excluded grade repeaters so that the typical-peer scores for all students in a grade were based on a pre-test score from the previous grade in the previous year. Doing so allows for meaningful comparisons between teachers, although at the cost of excluding some students from the value-added model. In addition to pre- and post-test scores, we collected data on other background characteristics of students, such as limited English proficiency status and poverty status.

The analysis file also contains a measure of the amount of instructional time each student spent with each teacher, which we refer to as dosage. Dosage enables us to assign teachers the appropriate amount of credit for each student based on two factors: (1) how much of the school year the student was in the teacher's class and (2) how much time the student spent with the teacher while enrolled. Some teachers participated in a pilot of a roster verification process in which they indicated whether and for how long they taught the students listed on their administrative rosters during each month of the school year. For teachers who participated in the pilot, we used these data to create a dosage for each teacher-student pair. However, most teachers did not teach in schools that participated in the pilot. For these teachers, we used school enrollment data to allocate proportional credit based on the fraction of time the student spent at the teacher's school.

We estimated the value-added model using four steps, each of which addressed a different conceptual challenge.

1. **Estimating a multiple regression model.** We used multiple regression, a statistical technique that enabled us to simultaneously account for a group of background factors to avoid holding teachers accountable for factors outside their control. We accounted for a set of student characteristics that could be related to performance on the OCCT or EOI post-tests. These characteristics include pre-tests in the same content area as the post-test, pre-tests in other content areas, poverty status, gender, race/ethnicity, existence of an individualized education plan, limited English language proficiency status, transfers of students across schools during the current (2012–2013) school year, and proportion of days the student attended school during the previous (2011–2012) school year. For OCCT post-

test scores in math and reading, we estimated models separately for each subject and grade. For the EOI post-test scores, we pooled eligible grades and estimated one model for each of the four subjects.

We weighted each student's contribution to a teacher's score by the proportion of time the student was assigned to the teacher when the teacher was teaching that subject. We used the Full Roster Method for teachers who shared students (Hock and Isenberg 2012). In some cases, a student was taught by one teacher for part of the year and another teacher for the rest of the year. In other cases, two or more teachers were jointly responsible for some of the same students at the same time. Using the Full Roster Method, teachers of shared students received equal credit for the students' achievement when the amount of instructional time was equal.

2. **Accounting for measurement error in the pre-test.** Because a student's performance on a single test is an imperfect measure of ability, teachers can be unfairly held accountable for the initial performance of their students, rather than being assessed on the gains they produce in student learning. Good or bad luck on the pre-test can dampen the observed relationship between pre- and post-test scores, compared to the true relationship between student achievement at the beginning and end of the year. If we were to use the observed relationships without any adjustments, teachers of students with low pre-test scores might be held partly accountable for the performance of their students before they entered their classrooms. To correct for this problem, we compensated for good or bad luck in pre-test scores—also known as measurement error—by employing a statistical technique that uses data on the reliability of each OCCT and EOI test provided by the test developers.

3. **Comparing teachers across grades.** The OCCT tests are not designed to allow the comparison of scores across grades. We therefore placed teachers on a common scale by translating each teacher's value-added estimate into a metric of generalized OCCT points. We based this translation on a three-stage procedure. First, before the multiple-regression step, we translated student test scores into a common metric in which each student test score is measured relative to other test scores within the same year, grade, and subject. In doing so, we set the average student test score to zero within each year, grade, and subject. We then used these scores to produce initial teacher value-added estimates. Second, we adjusted these estimates so that the average teacher in each grade and subject received the same estimate.[1] Third, we multiplied the resulting estimates by a grade-specific conversion factor to ensure that the dispersion of the estimates was similar by grade. For teachers with students in more than one grade, we took a student-weighted average of their grade-specific value-added results.

   Because the same EOI tests are given to students regardless of grade, we did not need to apply the same grade-level adjustments to the EOI value-added estimates. Instead, we accounted for grade as an additional student characteristic in the multiple regression model. Although not related to grade, we applied adjustments to EOI test scores and initial value-

[1] Although we mean-centered the student test scores and the other student characteristics in the regression, the initial value-added estimates also have to be mean-centered to account for differences in the weighting of the average due to different numbers of students contributing to each teacher's estimate.

added estimates similar to those for the OCCT model: (1) we translated student test scores into a common metric in which each student test score is measured relative to other test scores within the same year and subject and (2) we adjusted these estimates so that the average teacher in each subject received the same estimate.

4. **Accounting for imprecisely estimated measures based on few students.** Value-added estimates can be misleading if they are based on too few students. Some students might score well due to good luck rather than good knowledge of the material. For teachers with many students, good and back luck affecting test performance tends to cancel out. However, a teacher with few students can receive high or low value-added results due to luck. We made two adjustments to reduce this risk: (1) we reported estimates only for teachers with at least 10 students and (2) we used a statistical technique called shrinkage that accounts for the precision of the initial value-added estimate by combining the adjusted value-added estimate (from step 3) with the overall average teacher value-added estimate to produce a final value-added result (Morris 1983). Whereas shrinkage adjusts estimates for teachers with fewer students more toward the overall average, the adjustment is smaller for teachers with many students. Thus, we relied more heavily on an assumption of average effectiveness for teachers with few students.

C. Limitations of value-added models

All measures of teacher effectiveness have limitations, including those generated by value-added models. We discuss the following three limitations that are important to consider when using the results of value-added models.

1. **Estimation error.** Even with adjustments such as those described previously, value-added models generate estimates of teacher effectiveness that exhibit some uncertainty. As with any statistical model, there is uncertainty in the estimates produced due to limitations in the available data—such as the number of students taught by a teacher—and because of the methods used to generate the estimates. This uncertainty means that two teachers with similar value-added results could be statistically indistinguishable from each other, although one could in fact be more effective. To address this limitation, we have quantified the precision of the value-added results by reporting a confidence interval that gives a range of plausible value-added results for each teacher.
2. **Unmeasured differences between students.** Although value-added models account for differences in student performance based on documented student characteristics, such as poverty status and prior achievement, they cannot account for differences in student achievement that arise from unmeasured sources. Excluding these unmeasured student characteristics can lead to concerns about the interpretation of the results. For example, if some teachers are systematically paired with hard-to-teach students, those teachers might be penalized if the model does not account for the unobserved factors that cause the students to be difficult to teach. A related concern is that teachers' value-added results might reflect the efficacy of school inputs, such as principal leadership. To address this limitation, we note that experimental and quasi-experimental empirical studies suggest that these factors do not play a large role in determining teacher value added (Kane and Staiger 2008; Chetty et al. 2011; Kane et al. 2013).

3. **Single versus multiple measures.** Value-added results might not measure all aspects of teachers' contributions to students' learning or their school communities that are relevant to a comprehensive performance evaluation. Value-added results measure the contributions of teachers to their students' achievement on standardized test scores. Additional measures of teacher performance might improve the predictive power of teacher evaluation systems (Kane and Staiger 2012) or the future effectiveness of teachers (Taylor and Tyler 2011). Accordingly, OSDE plans that under full implementation of the TLE system, the overall TLE score will combine multiple measures of teacher performance.

II. KEY MODELING DECISIONS MADE BY THE STATE BOARD OF EDUCATION

We worked closely with the TLE Commission to design the value-added model. We identified the key decisions for the value-added model and then made recommendations for each of those decisions. The primary goal when forming our recommendations was to maximize the accuracy of the value-added results. We defined accurate value-added results as those that reflect teachers' true contributions to students' academic growth.[2] Our secondary goals included (1) maximizing the precision of value-added results, so teachers' value-added results are not different from their true contributions due to random chance; (2) maximizing the number of teachers who would receive value-added results; (3) ensuring that educators could understand the value-added results; and (4) making the best use of the available data. When we presented options to the TLE Commission, we highlighted the extent to which each option aligned with these criteria. In some instances, our recommendations balanced the trade-off between competing goals.

The TLE Commission made recommendations to the state board, which then made the final decisions about the design of the value-added model based on the TLE Commission's recommendations. Next, we describe how we implemented the key decisions for the value-added model and the rationale behind those decisions.

A. Determining which test scores to use in the value-added model

The first set of decisions pertains to which test scores to include in the value-added model. These decisions have important implications for which teachers get value-added results and which students are included in the creation of those results. The decisions apply to two types of tests: (1) pre-tests are assessments that took place in 2011–2012 in the previous grade and (2) post-tests are assessments that took place in 2012–2013. The state board approved four rules to define the pre- and post-test scores included in the value-added model.

Estimate value added using post-test scores in a subject only if pre-test scores in the same content area are available. We estimated value added based on the following subjects and grades by content area:

- **Math**: OCCT math in grades 4 through 8, algebra I in grades 8 and 9, algebra II in grades 9 through 11, and geometry in grades 9 through 12
- **Reading**: OCCT reading in grades 4 through 8 and English III in grade 11

We selected post-test scores in these subjects and grades because the students with these scores are most likely to have pre-tests in the same content area. We then excluded a student's post-test score if the student did not have a pre-test score in the same content area as the post-test. Pre-test scores from the same content area are typically the most important element used to estimate typical-peer scores because they have the strongest relationship with achievement on the post-test. As a result, the decision to require pre-test scores from the same content area improves the accuracy of the value-added results.

[2] Specifically, our goal was to achieve value-added results that minimize bias—those without systematic differences from teachers' true contributions.

Require that pre-test scores be from the previous grade in the previous year. We excluded students who did not have typical course-taking patterns from the value-added model. For example, we excluded students who were in the same grade in the 2011–2012 and 2012–2013 school years. A consequence of this decision is that value-added results will not reflect the achievement of students who repeat grades. However, value-added results that include these students who repeat grades might be inaccurate and imprecise because typical-peer scores for grade repeaters would be based on the achievement of the few students in the state who repeated their grade from the 2011–2012 school year.

Exclude Oklahoma Modified Alternate Assessment Program (OMAAP) and Oklahoma Alternate Assessment Program (OAAP) test scores from the value-added model. We excluded the OMAAP and OAAP test scores, which are from tests taken by students with diagnosed learning disabilities. The exclusion of these scores means that the value-added results will not reflect the contributions of teachers to the achievement of students who take these assessments. However, value-added results based on OMAAP and OAAP scores might be inaccurate and imprecise because scores on these assessments are not easily compared to scores on the OCCT and EOI assessments. In addition, relatively few students take the OMAAP and OAAP assessments. The consequences of this decision will be reduced over time because Oklahoma is phasing out the OMAAP assessments.

Account for pre-test scores in different content areas from the post-test. When available, we accounted for pre-test scores in the math, reading, and science content areas, regardless of the content area of the post-test. For example, value-added results based on grade 6 reading scores account for grade 5 math and science pre-test scores in addition to grade 5 reading scores. Because pre-test scores in more subjects could provide a more complete picture of a student's baseline level of academic achievement, accounting for pre-test scores in multiple content areas could improve the accuracy and precision of the value-added results.

B. Determining which other characteristics are in the model

Next, the state board defined which student background characteristics to include in the model. In addition to pre-test scores, we accounted for poverty status, gender, race/ethnicity, existence of an individualized education plan, limited English language proficiency status, transfers between schools during the school year, and school attendance during the prior school year. Doing so avoids holding teachers accountable for these factors that are outside of their control so that the value-added results more accurately reflect the contributions of teachers to students' academic growth.

C. Addressing the imprecision of the value-added results

Finally, the state board approved a rule to address the imprecision of the value-added results because imprecise results are unlikely to be useful to educators. The precision of a value-added result measures the risk that the result differs from the teacher's true contribution due to random chance.

To prevent using the most imprecise value-added results in the TLE system, we reported results only for teachers who had at least 10 students in the subject. Value-added results based on fewer than 10 students can be very imprecise and might not provide useful information to teachers about their effectiveness. Additionally, for teachers with at least 10 students, we used shrinkage to reduce the risk that they receive high or low value-added results due to luck.

III. CONSTRUCTION OF THE ANALYSIS FILE

In this chapter, we review the construction of the analysis file for the value-added model. First, we discuss the test scores used in the value-added model. We then discuss the data on student background characteristics used in the model. Finally, we discuss how we calculated teachers' shares of instruction for students with multiple teachers within each subject.

A. OCCT and EOI test scores

To be included in the value-added model, students' test score records must meet certain conditions based on when they were tested and whether we had a record of where the students were enrolled in school. The first set of conditions varies by test type:

- Students enrolled in grades 4 through 8 during the 2012–2013 school year were eligible to be included if they had an OCCT math or reading post-test score.
- Students with EOI scores in algebra I, geometry, algebra II, or English III from the 2012–2013 school year were eligible to be included if they were enrolled in a typical grade for the subject area. The typical grades were 8 and 9 for algebra I, grades 9 through 11 for geometry, grades 9 through 12 for algebra II, and grade 11 for English III.

The first two columns of Table III.1 summarize the post-test subjects and grades included in the value-added model.

Table III.1. Value-added model test subjects and grades

Post-test subject	Post-test grades	Same-content pre-test subjects
OCCT math	4 through 8	OCCT math
OCCT reading	4 through 8	OCCT reading
Algebra I EOI	8 and 9	OCCT math
Geometry EOI	9 through 11	OCCT math, algebra I, algebra II
Algebra II EOI	9 through 12	Algebra I, geometry
English III EOI	11	English II

Note: For a post-test score to be included in the value-added model, the student must have a pre-test score from the same content area in the previous grade.

We excluded test scores from the OMAAP and OAAP assessments from the analysis file. We then excluded students with post-tests from the analysis file for four reasons. First, we excluded conflicting post-test score records for the same test type or subject. Second, we required that students have a pre-test score in the same content area. For the OCCT post-tests, we required that students have pre-test scores in the same subject (math or reading). For EOI post-tests in algebra I, geometry, and algebra II, the pre-test scores must be another score in the math content area, which includes algebra I, geometry, algebra II, and OCCT math. For EOI post-tests in English III, we required that students have pre-test scores in English II. The third column of Table III.1 lists the possible same-content pre-tests for each post-test subject. Third, we excluded students who repeated or skipped a grade between the 2011–2012 and 2012–2013 school years. Finally, we dropped students from the analysis file if we they were not linked to a teacher eligible to receive a value-added result for the student's grade level during the 2012–2013 school

year. This occurred when a student was claimed only by a teacher with fewer than five students in his or her grade for OCCT post-tests or with fewer than five students overall for EOI post-tests (as we do not estimate a value-added measure for teachers with so few students).

Table III.2 shows the reasons students were excluded from the analysis files and the total numbers of students included in the models. The top row shows the total number of students with post-test scores. These counts represent students who could have been included in the analyses based only on having a post-test. The second through fifth columns show the totals for students in OCCT math and reading, and the last two columns show the totals for all four EOI subjects combined. As shown in the bottom row of the table, 90.5 percent of students with test scores from 2012–2013 were included in the analysis file for OCCT math, 90.3 percent for reading, and 85.2 percent for EOI subjects. The most common reason students were excluded was lack of a pre-test score in the same content area as the post-test score.

Table III.2. Reasons that students tested in 2013 were excluded from the analysis files

	OCCT math		OCCT reading		EOI subjects	
	Number	Percentage	Number	Percentage	Number	Percentage
Students with post-test scores	218,469	100.0	217,982	100.0	139,070	100.0
(1) Conflicting post-test scores	12	0.0	23	0.0	54	0.0
(2) Missing pre-test score from same content area	17,308	7.9	17,898	8.2	18,068	13.0
(3) Skipped or repeated a grade	1,571	0.7	1,562	0.7	1,136	0.8
(4) Not linked to an eligible teacher	1,832	0.8	1,606	0.7	1,317	0.9
Total excluded	**20,723**	**9.5**	**21,089**	**9.7**	**20,575**	**14.8**
Total included	**197,746**	**90.5**	**196,893**	**90.3**	**118,495**	**85.2**

Source: OSDE administrative data.

Notes: The table does not include 46,188 student-subject combinations who had OMAAP or OAAP scores from 2013, but no OCCT or EOI tests. Also, the 6,727 student-subject combinations taking an EOI test for the second time are excluded from the table.

Students are excluded sequentially in the order presented and so do not count for more than one reason in this table.

The columns for math and reading include students in grades 4 through 8. The EOI subjects are algebra I, geometry, algebra II, and English III.

For OCCT math, algebra I, geometry, or algebra II, the same-content pre-test score is another mathematics assessment. For OCCT reading or English III, the same-content pre-test score is another reading/ELA assessment.

For OCCT subjects, teachers must be linked to at least five eligible students in a single grade level to be considered eligible. For EOI subjects, teachers must be linked to at least five eligible students in any grade to be considered eligible.

Post-tested students not linked to any teacher in a subject are linked to catch-all teachers of unassigned students for the school-grade combination. These catch-all teachers are considered eligible teachers if they are linked to at least five eligible students. Across all grades and subjects, 29.4 percent of student-teacher links are to catch-all teachers of unassigned students.

The pre-test subjects associated with each EOI post-test subject vary across students because not all students take the EOI test in a subject in the same grade, and because there is no set order

that students must take the courses associated with the tests. We show the distribution of eligible pre-tests for each EOI post-test in Table III.3. For example, 73.9 percent of students with algebra I post-test scores in the analysis file have grade 8 OCCT math pre-tests, and 26.1 percent have grade 7 OCCT math pre-tests. These percentages sum to 100 percent because we required that all students in the analysis file have a pre-test score in the same content area. In contrast, only 99.2 percent of students with an algebra I post-test score have a pre-test score in the reading/ELA content area because we do not require that algebra I students have a reading/ELA pre-test.

Table III.3. Pre-test subjects of students by EOI post-test subject

	Post-test subject							
	Algebra I		Geometry		Algebra II		English III	
Pre-test subject	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
Math								
Grade 7 OCCT	8,277	26.1	0	0.0	0	0.0	0	0.0
Grade 8 OCCT	23,415	73.9	251	0.9	0	0.0	0	0.0
Algebra I	0	0.0	26,013	89.0	3,997	14.8	2,133	7.0
Geometry	0	0.0	0	0.0	22,978	85.2	16,741	54.7
Algebra II	0	0.0	2,972	10.2	0	0.0	9,171	30.0
Total	**31,692**	**100.0**	**29,236**	**100.0**	**26,975**	**100.0**	**28,045**	**91.7**
Reading/ELA								
Grade 7 OCCT	8,257	26.1	0	0.0	0	0.0	0	0.0
Grace 8 OCCT	23,175	73.1	6,908	23.6	0	0.0	0	0.0
English II	0	0.0	3,603	12.3	13,768	51.0	30,592	100.0
Total	**31,432**	**99.2**	**10,511**	**36.0**	**13,768**	**51.0**	**30,592**	**100.0**
Science								
Grade 8 OCCT	23,207	73.2	6,773	23.2	0	0.0	0	0.0
Biology I	0	0.0	9,384	32.1	13,945	51.7	16,560	54.1
Total	**23,207**	**73.2**	**16,157**	**55.3**	**13,945**	**51.7**	**16,560**	**54.1**

Source: OSDE administrative data.

The OCCT and EOI scores range from 400 to 999. However, scores on these scales are not designed to be meaningfully compared between grades, years, and subjects. To compare test scores across grades within each subject and year, we transformed the test scores in a two-part process. First, we subtracted the mean score and divided by the standard deviation for each grade, subject, and year to obtain a *z*-score.[3] This step translated scores in every grade and subject into a common metric. Second, we created a measure with a range resembling the original test score scale by multiplying each *z*-score by a common factor across all grades within each subject and year. The common factor was equal to the square root of the average variance across all grades for each subject and year.

B. Student background characteristics

We used the data provided by OSDE to construct variables used as controls for student background characteristics in the value-added model. The value-added model accounts for the following:

- Pre-test in the same content area as the post-test
- Pre-tests in other content areas (including math, reading/ELA, and science, when available)
- Poverty status
- Gender
- Race/ethnicity
- Existence of an individualized education plan
- Limited English language proficiency status
- Transfers of students across schools during the 2012–2013 school year
- Proportion of days the student attended school during the 2011–2012 school year

Attendance is a measure of student motivation. We used previous—rather than current-year—attendance to avoid confounding student attendance with current-year teacher effectiveness; that is, a good teacher might be expected to motivate students to attend school more regularly than a weaker teacher would. The proportion of the days a student attended school is a continuous variable that could range from zero to one. Because some districts did not provide OSDE with student-level attendance records from the 2011–2012 school year, we used the student's schoolwide attendance rate in place of the student's individual attendance rate for 21.9 percent of students in the analysis files.

We accounted for whether a student transferred into or out of a teacher's school during the school year because students who transferred schools might have experienced a disruptive environment outside of school that led to lower performance compared to similar students who did not transfer between schools. By accounting for mid-year transfers, teachers of transfer

[3] Subtracting the mean score for each subject and grade creates a score with a mean of zero in all subject-grade combinations.

students were not held accountable for circumstances outside their control that were related to these disruptions.

Aside from attendance and pre-test variables, the student background variables are binary, taking a value of zero or one. Table III.4 shows the characteristics of students in the OCCT and EOI analysis files.

Table III.4. Characteristics of students from the 2012–2013 school year

	OCCT math		OCCT reading		EOI subjects	
Characteristic	Number	Percentage	Number	Percentage	Number	Percentage
Included in the value-added model	197,746	100.0	196,893	100.0	118,495	100.0
Eligible for free lunch	86,571	43.8	86,554	44.0	68,256	57.6
Eligible for reduced-price lunch	18,189	9.2	18,069	9.2	10,267	8.7
Female	99,524	50.3	99,681	50.6	60,835	51.3
African American	21,925	11.1	22,009	11.2	11,693	9.9
Hispanic	27,467	13.9	27,107	13.8	13,232	11.2
American Indian	45,295	22.9	45,162	22.9	24,570	20.7
Asian/Pacific Islander	5,724	2.9	5,646	2.9	4,192	3.5
Individualized education plan with accommodations	9,156	4.6	7,745	3.9	3,998	3.4
Individualized education plan without accommodations	6,640	3.4	6,513	3.3	1,694	1.4
Limited English proficient with accommodations	3,495	1.8	3,163	1.6	1,499	1.3
Limited English proficient without accommodations	6,106	3.1	5,879	3.0	1,079	0.9
Transferred schools during the school year	12,475	6.3	12,480	6.3	4,659	3.9

Source: OSDE administrative data.

Notes: All percentages are based on the counts in the top row.

Student characteristics were calculated as a weighted average for students enrolled in multiple schools during the year. The counts and percentages were not weighted in any other way.

For all student characteristics in this table, less than 2 percent of students have missing data.

We imputed data for students included in the analysis file but who had missing values for one or more student characteristics. Our imputation approach used the values of nonmissing student characteristics to predict the value of the missing characteristic. Less than 4 percent of students in the value-added analysis files had any characteristic imputed. Most imputed values were for missing pre-test scores in different content areas from the OCCT post-test.[4] We did not generate imputed values for the same-content pre-test; we dropped from the analysis file any students with missing same-content pre-test scores. Finally, we did not impute any missing pre-test scores for students in the EOI analysis file.

[4] In addition to imputing values for some of the characteristics included in Table III.4, we also generated imputed values of attendance during the 2011–2012 school year for approximately 3 percent of students.

C. Teacher dosage

Some students were taught by a team of teachers, either because they moved between schools or were taught by multiple teachers in a school. We refer to the fraction of the time that a student was enrolled with a given teacher for a subject as the dosage. In this section, we describe how we calculated dosage for the value-added model.

1. Roster verification

In the 2012–2013 school year, OSDE implemented a pilot roster verification program in selected schools.[5] Roster verification is a process by which records of teachers' monthly shares of instruction for each student and course are submitted and either verified or corrected by teachers and school administrators. For example, consider a student who spends 2.5 days per week in teacher A's classroom learning math and 2.5 days per week in teacher B's classroom learning math. This student would be recorded as having spent 50 percent of math instructional time with teacher A for that month. Likewise, the same student would also be recorded as having spent 50 percent of math instructional time with teacher B for that month. In recording the share of instructional time with a student, teachers rounded to the nearest quarter. Thus, 0, 25, 50, 75, and 100 percent were the possible responses, representing the categories of none, some, equally shared, most, and all. After these shares were reported, they were verified or corrected by those teachers and school administrators. The roster verification process differed slightly in the Tulsa public schools, where teachers instead rounded to the nearest 10 percent. Roster verification was not implemented statewide and was not always fully implemented in the pilot schools.

2. Dosage

Teacher dosage measures the proportion of instructional time a teacher spent with a student during the school year. Because most students take the post-test in April or early May, we excluded instructional time in May from the calculation of dosage. To calculate teacher dosage for a student-teacher link, we used a three step process: (1) we determined the amount of instructional time in each month of the school year, (2) we summed the monthly dosages, and (3) we divided by the number of months to obtain dosage as a percentage of the school year through April. In doing so, we summed monthly dosages for a student-teacher combination when the student was linked to the teacher in multiple courses in the same subject. Thus a student-teacher link could have a dosage that exceeds 100 percent. For example, a teacher who claimed the same student in roster verification for two full-year courses, and assigned 100 percent instructional time in every month for both course would have a combined dosage of 200 percent for that student. When two or more teachers claimed the same students at 100 percent during the same term, we assigned each teacher full credit for the shared students. Thus solo-taught and co-taught students contribute equally to teachers' value-added results.

Although we obtained monthly dosage from the roster-verified records when possible, roster verification was not implemented statewide. For teachers without verified roster data, we assumed that the monthly dosage for a teacher-student link was equal to the proportion of instructional days the student was officially enrolled in the school based on administrative data

[5] The full roll-out of roster verification for Oklahoma is planned for the 2014–2015 school year.

from OSDE. These data contained dates of school withdrawal and admission, as well as school calendars. We assume that learning accumulated at a constant rate and, therefore, treat days spent at one school as interchangeable with days spent at another. For example, if a student split time equally between two schools, we set the dosage at each school to 50 percent, regardless of which school the student first attended.

3. Catch-all teachers

Some students appear to not be linked to a math or reading/ELA teacher because roster verification was not implemented statewide and because of limitations of the administrative data linking teachers to students provided by OSDE. For example, some students with post-test scores in the analysis file were not linked to a teacher in the subject of the post-test. All such students were assigned a placeholder teacher for each subject in which they had no roster record linking them to a teacher. We created a so-called catch-all teacher in OCCT subjects for each school-grade combination that had unlinked post-tested students. We did the same in EOI subjects for each school that had unlinked post-tested students. Teacher dosages were assigned to catch-all teachers in the same manner as teachers with unverified roster records.

IV. TECHNICAL DETAILS OF THE VALUE-ADDED MODEL

In this chapter, we describe the technical details of the value-added model. We organize the discussion into four topics: (1) the estimation equations, (2) how we address measurement error in the pre-tests, (3) how we generalize estimates to be comparable across grades, and (4) how we account for the number of students per teacher.

A. Estimation equations

We developed two linear regression approaches based on whether the post-test score was from the OCCT or EOI. For the OCCT approach, we estimated regression models separately for each grade and subject combination. This approach covered grades 4 through 8 and the subjects reading and math. For the EOI approach, we estimated the regression models separately for each subject only. The subjects were algebra I, geometry, algebra II, and English III. For these EOI subjects, we pooled the regression models by subject across grades because the assessment for a given subject is the same in any grade that it is taken. For each EOI subject, we included in the model only the grades in which most students had post-test scores. The grades were grades 8 and 9 for algebra I, grades 9 through 11 for geometry, grades 9 through 12 for algebra II, and grade 11 for English III.

For both the OCCT and EOI approaches, the post-test score depends on pre-test scores, student background characteristics, the student's teacher, and unmeasured factors. For a given teacher *t* and student *i*, in grade *g*, the regression equation for the OCCT approach is:

$$(1)\quad Y_{tig} = \lambda_{Mg} M_{i(g-1)} + \lambda_{Rg} R_{i(g-1)} + \lambda_{Sg} S_{i(g-1)} + \boldsymbol{\beta}'_{1g} \mathbf{X}_i + \boldsymbol{\delta}'_{1g} \mathbf{T}_{1tig} + \varepsilon_{1tig},$$

and the regression equation for the EOI approach is:

$$(2)\quad Y_{tig} = \boldsymbol{\lambda}'_2 \mathbf{P}_{i(g-1)} + \boldsymbol{\gamma}' \mathbf{C}_{i(g-1)} + \boldsymbol{\kappa}' \mathbf{G}_g + \boldsymbol{\beta}'_2 \mathbf{X}_i + \boldsymbol{\delta}'_2 \mathbf{T}_{2ti} + \varepsilon_{2tig},$$

In both regression equations, Y_{tig} is the post-test score. The regressions are run separately by post-test subject, so we drop a subject subscript for ease of notation. In the OCCT equation, $M_{i(g-1)}$ is the math pre-test for student *i* from the previous grade *g-1*, and $R_{i(g-1)}$ is the reading pre-test from the previous grade. For 6th-grade students, we also include $S_{i(g-1)}$, the science pre-test taken in the previous grade.

In the EOI equation, the vector $\mathbf{P}_{i(g-1)}$ denotes variables for pre-test scores in each of the included subjects in the previous grade. Unlike the OCCT approach, in the EOI approach not all students with a post-test in a given subject have scores from the same pre-tests. For example, the regression for the geometry post-test included five pre-test variables—grade 8 math, algebra I, algebra II, English II, and biology. Because all students included in the EOI models were required to have a pre-test score on a same-content assessment, students included in the regression for the geometry post-test have a pre-test score in one of grade 8 math, algebra I, or algebra II. Some students in the geometry regression additionally have pre-test scores in English II and/or biology, but these were not required for all students. To account for different pre-test subjects, the EOI equation includes binary variables in $\mathbf{C}_{i(g-1)}$ that indicate whether a student had a pre-test in each subject.

The pre-test scores in both equations capture prior inputs into student achievement; we estimated the associated coefficients—λ_{Mg}, λ_{Rg}, λ_{Sg}—and the vector $\boldsymbol{\lambda}_2$, using a procedure that corrects for measurement error in these pre-test scores. The subscript 2 distinguishes the EOI model coefficients from the OCCT model coefficients.

The vector $\mathbf{X}_i$ denotes the control variables for student background characteristics. For the OCCT approach, we allowed these coefficients to vary by grade, represented in Equation (1) by the g subscript on the vector $\boldsymbol{\beta}_{1g}$. Because there are some grade and subject combinations for EOI post-test scores with very few students, we did not allow the coefficients to vary by grade for the EOI approach. Doing so leads to more precise coefficient estimates in $\boldsymbol{\beta}_2$ and could lead to more precise value-added results.

The vectors $\mathbf{T}_{1tig}$ and $\mathbf{T}_{2ti}$ consist of binary variable for each teacher. Because the OCCT approach is estimated separately by grade, a teacher who taught multiple grades had variables in each grade regression model. For example, a teacher who taught math in grades 4 and 5 had one variable in $\mathbf{T}_{1ti4}$ for the grade 4 regression and one in $\mathbf{T}_{1ti5}$ for the grade 5 regression. The EOI approach is not grade-specific, so a teacher who taught multiple grades had only one variable in $\mathbf{T}_{2ti}$. Each teacher-student observation has one nonzero element in $\mathbf{T}_{1tig}$ and/or $\mathbf{T}_{2ti}$. The coefficient vectors $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ contain the initial value-added estimates. Rather than dropping one element of $\mathbf{T}_{1tig}$ or $\mathbf{T}_{2\text{ti}}$ from the regression, we estimated the regression models without constant terms. The vectors $\mathbf{T}_{1tig}$ and $\mathbf{T}_{2ti}$ also include binary variables for the catch-all teachers of unassigned students. We also mean-centered the control variables so that each element of $\boldsymbol{\delta}_{1g}$ and $\boldsymbol{\delta}_2$ represents a teacher-specific intercept term for a student with average characteristics.

Because we estimated the model separately by grade, the OCCT approach does not include grade variables. However, we adjusted for factors beyond teachers' control that might drive cross-grade differences in value added using the approach described in Section C. For the EOI approach, we included binary variables for each grade in the vector $\mathbf{G}_g$. We excluded the indicator for the highest included grade for each subject. These variables account for post-test score variation by grade that are due to differences in students' ability that are not accounted for in the model.

Table IV.1 shows the coefficient estimates and standard errors of the control variables in the model by subject and grade span. The top panel shows the average association between the pre-tests and achievement on the post-tests (measured in points on the test), accounting for all other variables in the regression. The bottom panel shows the association between each student characteristic and post-test scores.

To account for team teaching, we used the Full Roster Method (Hock and Isenberg 2012). In this approach, each student contributes one observation to the model for each teacher to whom he or she was linked. Thus, the unit of observation in the analysis file is a teacher-student combination. This method is based on the assumption that teachers contribute equally to student achievement within each team. We weighted each record based on the dosage associated with the teacher-student combination using weighted least squares (WLS). We used a cluster-robust sandwich variance estimator (Liang and Zeger 1986; Arellano 1987) to address the correlation in the error terms and heteroskedasticity. The correlation arises from the presence of students with multiple records in each regression model. The heteroskedasticity arises from differences across

students in how well the model can predict post-test scores based on the background characteristics included in the regressions. Our approach leads to standard errors that are consistent in the presence of both heteroskedasticity and clustering at the student level.

The regression models yield initial value-added estimates for each grade and subject for the OCCT approach and for each subject in the EOI approach. For the OCCT approach, we included teachers in the regression model only if they had at least five students in that grade and subject combination. For the EOI approach, we estimated initial value-added estimates for teachers with at least five students across all eligible grades.

How to Interpret Table IV.1

Table IV.1 displays the regression coefficients from the value-added model. In other words, it describes the relationships between the characteristics of Oklahoma students and achievement on the post-test, accounting for all other variables in the regression. The coefficients give the amount of the increase—or decrease if the coefficient is negative—in the typical-peer score when a characteristic increases by one unit. For example, the coefficient of 0.69 in the first row of the first column of the table indicates that an increase by one OCCT point on a student's math pre-test score is associated with a 0.69 point increase in the student's typical-peer score on the OCCT math post-test. Similarly, the coefficient on the fraction of the prior year a student attended school indicates that a student who attended 100 percent of the prior year is predicted to score 34.03 points higher than the prediction if the student instead had attended for none of the prior year. More than 99 percent of students attended 75 percent of the prior year or more, so the typical contribution of prior attendance to the typical-peer score is much smaller than this change of 34.03 points might suggest; the change in typical-peer scores associated with a change in attendance from 75 to 100 percent is 8.51 OCCT points.

For characteristics that are binary indicators, the coefficient gives the increase in the predicted score for a student who has that characteristic relative to a student who does not. For example, students in grades 4 through 8 math who transferred between schools during the school year are predicted to score 7.94 points lower than students who did transfer. The other binary indicators are in groups of related indicators. For example, the coefficients on the two indicators of student poverty status measure the difference in the predicted score of a student with that status (for example, students eligible for reduced-price lunch) relative to a student who is eligible for free lunch.

Each regression coefficient describes a relationship after accounting for all other characteristics included in the model. Put another way, the coefficient on a characteristic gives the change in predicted achievement when the characteristic is changed from no to yes or increased by one point, assuming that all of the students' other characteristics remain the same. Consequently, coefficients might not reflect the relationship we would observe had the other characteristics not been accounted for in the value-added model. This feature of multiple regression coefficients can produce counterintuitive relationships between characteristics and achievement if the contributions of one characteristic are accounted for largely by another characteristic in the model. For example, coefficients on limited English proficiency status would likely be consistently negative and greater in magnitude if the model did not also account for students' pre-test scores, because students with limited English proficiency tend to have lower pre-test scores.

To put the magnitude of the coefficients into perspective, they can be compared to the typical range of student achievement on an OCCT or EOI test. The standard deviation of student achievement on the grade 4 math post-test was 87.7 OCCT points, indicating that about two-thirds of students scored within 87.7 points above or below the average score on the assessment. The standard deviations for other grades ranged from 75.8 to 87.7 points in grades 4 through 8 math, from 64.5 to 80.4 points in grades 4 through 8 reading, and from 48.0 to 75.9 points in the four EOI subjects.

The number in parentheses below each coefficient is the standard error of the coefficient—a measure of precision. A more precise coefficient indicates with more certainty that a coefficient reflects the actual relationship between the characteristic and achievement. Coefficients with smaller standard errors are more precise. The coefficients on the pre-tests are more precise than those on individual background characteristics. Roughly, a coefficient that is at least twice as large as its standard error is said to be statistically significant, meaning that it is likely that the direction of the relationship—whether positive or negative—reflects the actual relationship between the characteristic and achievement and is unlikely to be produced by chance.

Table IV.1. Coefficients on covariates in the value-added models, by post-test subject

	Grades 4 through 8 OCCT		EOI subjects			
Variable	**Math**	**Reading**	**Algebra I**	**Geometry**	**Algebra II**	**English III**
Pre-test scores (average coefficients across grades for OCCT post-tests and subjects for EOI post-tests)						
Math	0.69	0.13	0.40	0.74	0.71	0.07
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Reading/ELA	0.14	0.63	0.07	0.04	0.04	0.56
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)
Science	0.12	0.29	0.17	0.19	0.15	0.05
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Individual student background characteristics (average coefficients across grades for OCCT post-tests)						
Ineligible for free or reduced-price lunch	5.37	5.82	1.06	2.89	1.94	1.47
	(0.69)	(0.68)	(0.44)	(0.60)	(0.79)	(0.46)
Eligible for reduced-price lunch	2.13	2.68	-0.37	2.67	-1.05	1.20
	(1.00)	(1.00)	(0.66)	(0.90)	(1.17)	(0.70)
Female	-2.31	6.75	3.87	-2.88	1.69	5.63
	(0.57)	(0.56)	(0.36)	(0.49)	(0.60)	(0.36)
African American	-2.92	-3.71	0.14	-7.28	3.28	-1.22
	(1.05)	(1.05)	(0.70)	(0.90)	(1.23)	(0.71)
Hispanic	-0.20	-0.86	2.52	-0.28	3.40	-0.04
	(0.99)	(0.97)	(0.67)	(0.88)	(1.10)	(0.66)
American Indian	-1.00	-0.80	-0.50	-2.54	-1.77	0.04
	(0.73)	(0.71)	(0.46)	(0.63)	(0.81)	(0.47)
Asian/Pacific Islander	8.44	2.60	10.60	5.69	11.45	0.83
	(1.78)	(1.80)	(1.36)	(1.58)	(1.71)	(1.11)
Individual education plan with accommodations	-9.10	-14.02	-3.13	-13.91	-18.02	-7.61
	(1.53)	(1.62)	(1.32)	(1.51)	(2.22)	(1.20)
Individual education plan without accommodations	-3.23	-5.02	0.19	-2.81	-4.22	-4.19
	(1.74)	(1.61)	(1.88)	(2.40)	(3.01)	(1.65)
Limited English proficiency with accommodations	-0.79	-12.89	3.20	-3.56	8.11	3.56
	(2.85)	(2.72)	(1.70)	(2.96)	(4.55)	(2.78)
Limited English proficiency without accommodations	0.32	-7.68	3.21	-6.10	1.76	6.39
	(1.85)	(1.85)	(1.80)	(3.11)	(4.69)	(2.63)
Transferred schools during the school year	-7.94	-4.17	-6.48	-6.29	-11.65	-4.24
	(1.33)	(1.27)	(1.03)	(1.39)	(2.30)	(1.11)
Fraction of the prior year student attended school	34.03	-3.02	12.92	52.25	75.03	8.50
	(20.04)	(19.41)	(12.27)	(16.00)	(23.58)	(10.63)

Source: Mathematica calculations based on OSDE administrative data.

Notes: Standard errors are in parentheses.

For OCCT post-tests, the reported coefficient estimates represent weighted averages of the coefficients estimated separately for each grade, where the weights are the number of student equivalents in the grade. Additionally, for EOI post-tests, the reported coefficient estimates of pre-test scores represent weighted averages of the coefficients estimated separately for each pre-test math, reading/ELA, or science subject, where the weights are the number of student equivalents with a pre-test score in a given subject. The associated standard errors similarly represent weighted averages across grades or subjects. The standard errors therefore do not account for the variability of the estimates across grades or subjects. These numbers are presented for descriptive purposes only and should not be used to conduct rigorous statistical tests.

The math pre-test score is the OCCT math assessment from the previous grade for students in the OCCT math value-added model. The math pre-test score for students in the EOI models is another math OCCT or EOI assessment from the previous grade. The case for reading is analogous to that for math. OCCT science pre-test scores are included for students in grades 6 and 9, and EOI biology pre-test scores are included for students who took the test in the previous grade.

The coefficients on poverty status variables are relative to students who are eligible for free lunch—the excluded category. Similarly, the coefficients on racial/ethnic variables are interpreted relative to students who are in the white or other racial and ethnic categories, as variables for both of these groups are excluded from the regression.

n.a. = not applicable.

B. Measurement error in the pre-tests

We corrected for measurement error in the pre-tests using data on test reliability. As a measure of true student ability, student achievement tests contain measurement error. This error causes standard models to produce biased initial value-added estimates. To address this serious issue, we implemented the errors-in-variables correction (Buonaccorsi 2010). The correction nets out the known amount of measurement error as measured by the reliability of the OCCT and EOI tests available from the test publisher (Pearson 2012a, 2012b).

In practice, to account for measurement error in the pre-tests, we estimated Equations (1) and (2) using two regression steps. In the first step, we accounted for measurement error. Because of computational limitations with the method used to account for measurement error, we could not obtain measures of the precision of value-added estimates from the first-stage regression. Thus, we needed a second step to calculate the precision of the estimates.

In the first regression step, we used the errors-in-variables approach to get the initial estimates of value added. We estimated the regression Equations (1) and (2) with the correction to obtain unbiased estimates of the coefficients on the pre-test scores. We based the correction on the published reliabilities for each OCCT and EOI test. For the OCCT model, we used grade- and subject-specific reliability data. For the EOI regression, we used subject-specific reliability data. We then used the coefficients to calculate an adjusted post-test score that nets out the contribution of the pre-test scores. The adjusted post-test score for the OCCT approach is given by:

$$(3) \quad A_{1tig} \equiv Y_{tig} - \lambda_{Mg} M_{i(g-1)} - \lambda_{Rg} R_{i(g-1)} - \lambda_{Sg} S_{i(g-1)}$$

The adjusted post-test scores for the EOI approach are given by:

$$(4) \quad A_{2tig} \equiv Y_{tig} - \boldsymbol{\lambda}_2' \mathbf{P}_{i(g-1)}$$

The vectors A_{1tig} and A_{2tig} represent the post-test scores, net of the estimated contribution of the student's pre-test scores. We calculated an adjusted post-test score for each OCCT grade and subject and for each EOI subject.

We used these adjusted post-tests scores in a second regression step to obtain standard errors that are consistent in the presence of both heteroskedasticity and clustering at the student level, because the regression includes multiple observations for the same student. This second-stage regression is necessary because it is not computationally possible to simultaneously account for correlation in the error term ε_{2tig} across multiple observations and apply the numerical formula for the errors-in-variables correction. Thus, for each OCCT grade and subject we estimated the final regression in Equation (5):

$$(5) \quad A_{1tig} = \boldsymbol{\beta}_{1g}' \mathbf{X}_i + \boldsymbol{\delta}_1' \mathbf{T}_{tig} + \varepsilon_{tig} .$$

and for each EOI subjects, we estimated the final regression in Equation (6):

(6) $A_{2tig} = \boldsymbol{\gamma}'\mathbf{C}_{i(g-1)} + \boldsymbol{\kappa}'\mathbf{G}_g + \boldsymbol{\beta}_2'\mathbf{X}_i + \boldsymbol{\delta}_2'\mathbf{T}_{2ti} + \varepsilon_{2tig}$.

The coefficients appear in Equations (5) and (6) as they did in Equations (1) and (2) because the regressions produce identical coefficient estimates; Equations (5) and (6) apply a correction only to the standard errors.

This two-step method likely underestimates the standard error of the estimated $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ because the adjusted gains in Equations (3) and (4) rely on the estimated values of λ_{Mg}, λ_{Rg}, λ_{Sg}, and the vector $\boldsymbol{\lambda}_{(g-1)}$. Treating these coefficients as fixed rather than as estimates does not fully account for variability in post-test scores related to pre-test scores. Nonetheless, with the large within-grade and within-subject sample sizes, the pre-test coefficients were precisely estimated, likely leading to a negligible difference between the robust and clustering-corrected standard errors.

Underestimated standard errors could result in insufficient shrinkage of some teachers' value-added estimates, which we discuss in Section D. When using value-added point estimates for teacher evaluations, the key concern is not whether the standard errors of the estimates are universally underestimated, but whether the standard errors for some teachers are disproportionately underestimated, which can lead to some teacher estimates shrinking too little relative to other teacher estimates in the final step. Thus, there is a trade-off in the design of the model between insufficient shrinkage for some teachers and accounting for measurement error. This approach emphasizes accuracy and face validity of teachers' value-added estimates over any consequences of underestimated standard errors for the shrinkage procedure.

C. Generalizing estimates to be comparable across grades

Both the average and variability of value-added estimates can differ across grade levels, which can prevent meaningful comparison of teachers assigned to different grades. The main concern is that factors beyond teachers' control might drive cross-grade discrepancies in the distribution of value-added estimates. For example, the standard deviation of adjusted post-test scores might vary across grades as a consequence of differences in the alignment of tests or the retention of knowledge between years. However, in the TLE system, all teachers will be compared within a subject, regardless of any grade-specific factors that might affect the distribution of gains in student performance between years.

Because of differences in our approach to estimating value-added based on the OCCT and EOI tests, our method to address differences across grades also varied. For the OCCT approach, we transformed the grade-specific value-added estimates to be comparable across grades and then combined these transformed estimates for teachers of multiple grades. For the EOI approach, we addressed differences across grades by accounting for grade in the regression model.

1. Grade-level adjustments in the OCCT approach

Transforming estimates into generalized OCCT points. For value added based on the OCCT tests, we translated teachers' grade-level estimates so that each set of estimates is expressed in a common metric of generalized OCCT points. Aside from putting value-added estimates for teachers onto a common scale, this approach leads to distributions of teacher

estimates that are more equal across grades. Doing so avoids penalizing or rewarding teachers simply for teaching in a grade with atypical test properties. However, the approach does not reflect a priori knowledge that the true distribution of teacher effectiveness is similar across grades. Rather, without a way to distinguish cross-grade differences in teacher effectiveness from cross-grade differences in testing conditions, the test instrument itself, or student cohorts, the approach reflects an implicit assumption that the distribution of true teacher effectiveness is the same across grades.

We standardized the estimated regression coefficients from the OCCT regressions so that the mean and standard deviation of the distribution of teacher estimates is the same across grades. First, we subtracted from each unadjusted estimate the average of all estimates within the same grade. We then divided the result by an estimate of the standard deviation within the same grade. To reduce the influence of imprecise estimates obtained from teacher-grade combinations with few students, we calculated the average using weights based on the number of students taught by each teacher. Our method of calculating the standard deviation of teacher effects also assigns less weight to imprecise individual estimates. Finally, we multiplied by the square root of the teacher-weighted average of the grade-specific variances, obtaining a common measure of effectiveness on the generalized OCCT-point scale.

Formally, the value-added estimate expressed in generalized OCCT points is the following:

$$(7) \quad \hat{\eta}_{tg} = \frac{\left(\hat{\delta}_{tg} - \overline{\hat{\delta}_g}\right)}{\hat{\sigma}_g} \times \sqrt{\left(\frac{1}{K}\sum_{h=4}^{8} K_h \hat{\sigma}_h^2\right)},$$

where $\hat{\delta}_{tg}$ is the grade-*g* estimate for teacher *t*, $\overline{\hat{\delta}_g}$ is the weighted average estimate for all teachers in grade *g*, $\hat{\sigma}_g$ is the estimate of the standard deviation of teacher effectiveness in grade *g*, K_h is the number of teachers with students in grade *h*, and *K* is the total number of teachers.

In Equation (7), we used an adjusted standard deviation that removes estimation error to reflect the dispersion of underlying teacher effectiveness. The unadjusted standard deviation of the value-added estimates will tend to overstate the true variability of teacher effectiveness; because the scores are regression estimates, rather than known quantities, the standard deviation will partly reflect estimation error. Using the unadjusted standard deviations to scale estimates could lead to over- or underweighting one or more grades when the extent of estimation error differs across grades. This is because doing so would result in estimates with the same amount of total dispersion—the true variability of teacher effectiveness and the estimation error combined—in each grade, but the amount of true variability in each grade would not be equal. Instead, we scaled the estimates using the adjusted standard deviation so that estimates of teacher effectiveness in each grade have the same adjusted standard deviation by spreading out the distribution of effectiveness in grades with relatively imprecise estimates.[6]

[6] For teachers in grades with imprecise estimates, the shrinkage procedure, described in Section D, counteracts the tendency for these teachers to receive final estimates that are in the extremes of the distribution.

We calculated the error-adjusted variance of teacher value-added results separately for each grade as the difference between the weighted variance of the grade-*g* teacher estimates and the weighted average of the squared standard errors of the estimates. The error-adjusted standard deviation $\hat{\sigma}_g$ is the square root of this difference. We chose the weights based on the empirical Bayes (EB) approach outlined by Morris (1983).

Table IV.2 shows the adjusted standard deviation of the initial estimates of teacher effectiveness derived from the value-added regression as well as the weighted average across all grades produced by Equation (7). A higher standard deviation for a grade-year combination indicates more dispersion in underlying teacher effectiveness before the transformation into generalized OCCT points. The standard deviation of value-added estimates ranged from 18.7 to 19.8 OCCT points in math and from 11.0 to 12.1 points in reading. By comparison, the range of the standard deviations of student-level achievement across grades was 75.8 to 87.7 OCCT points in math and 64.5 to 80.4 points in reading. Because we estimated value-added results for EOI subjects pooling all grades, we report only the combined standard deviations. These ranged from 5.7 to 23.8 EOI points. By comparison, the range of the standard deviations of student-level achievement across grades was 48.0 to 75.9 EOI points.

Table IV.2. Student-weighted standard deviations of value-added results

Model	Grade					Weighted average	Student-level achievement
	4	5	6	7	8		
OCCT math	19.8	19.4	19.3	18.7	19.3	19.8	75.8 – 87.7
OCCT reading	12.1	11.6	11.7	11.0	11.5	11.8	64.5 – 80.4
Algebra I						15.4	48.0
Geometry						23.8	75.9
Algebra II						15.8	64.1
English III						5.7	48.8

Source: Mathematica calculations based on OSDE administrative data.

Combining OCCT estimates for teachers of multiple grades. To combine grade-level estimates from OCCT models into a single value-added result, denoted as $\hat{\eta}_t$, for a teacher with students in multiple grades, we used a weighted average of the grade-specific estimates (expressed in generalized OCCT points). We set the weight for grade *g* equal to the proportion of students of teacher *t* in grade *g*. Because combining teacher effects across grades might cause the overall average to be nonzero, we recentered the estimates on zero before proceeding to the next step.

We computed the variance of each teacher's combined effect as a weighted average of the grade-specific squared standard errors of the teacher's estimates. We set the weight for grade *g* equal to the squared proportion of students of teacher *t* in grade *g*. For simplicity, we assumed that the covariance across grades is zero. In addition, we did not account for uncertainty arising because $\overline{\hat{\delta}_g}$ and $\hat{\sigma}_g$ in Equation (7) are estimates of underlying parameters rather than known constants. Both decisions imply that the standard errors will be underestimated slightly.

2. Grade-level adjustments in the EOI approach

Unlike the grade-by-grade OCCT approach, we pooled grades for the EOI estimation equation and estimated a single initial value-added estimate for each teacher, rather than one for each teacher-grade combination. We also used a different method to account for differences across grades because of this difference in approach. To account for differences across grades in EOI models, we included binary variables for each grade (excluding the highest grade) in the regressions. For example, students taking geometry in grade 9 or 10 might be higher-ability students on average compared to students taking geometry in grade 11, even after accounting for the other variables in the model. The coefficients on the grade 9 and grade 10 indicators would give the increase in the typical achievement for students in these grades relative to a student taking geometry in grade 11. This approach avoids penalizing teachers for teaching in grades with lower-ability students.

One potential concern with this approach is that the coefficients could also measure differences in the effectiveness of teachers in different tracks, which could lead to bias in the value-added results related to teachers' grade assignments. This is unlikely to occur because the coefficients on the grade indicators are based on comparing the achievement of students of teachers with students in multiple grades for the same subject, instead of comparing achievement across teachers. We achieved this by simultaneously including $\mathbf{T}_{2ti}$ in the regression equation so that all relationships between the variables included in the regression and achievement—including the grade indicators—were based on within-teacher variation in student achievement, rather than on variation from students in different teachers' classrooms.

D. Accounting for different numbers of students

To reduce the risk that teachers, particularly those with relatively few students in their grade, will receive a very high or very low effectiveness measure by chance, we applied the EB shrinkage procedure (Herrmann et al. 2013). Using the EB procedure outlined in Morris (1983), we computed a weighted average of an estimate for the average teacher and the initial estimate based on each teacher's own students. For teachers with relatively imprecise initial estimates based on their own students, the EB method effectively produces an estimate based more on the average teacher. For teachers with more precise initial estimates based on their own students, the EB method puts less weight on the value for the average teacher and more weight on the value obtained from the teacher's own students.

The EB estimate for a teacher is approximately equal to a precision-weighted average of the teacher's initial estimated effect and the overall mean of all estimated teacher effects.[7] Following the standardization procedure, the overall mean is zero, with better-than-average teachers having positive scores and worse-than-average teachers having negative scores. We therefore arrived at the following:

[7] In Morris (1983), the EB estimate does not exactly equal the precision-weighted average of the teacher's initial estimated effect and the overall mean of all estimated teacher effects. This adjustment decreases the weight on the estimated effect by $(K - 3)/(K - 1)$, where K is the number of teachers. For ease of exposition, we have omitted this correction from the description given here.

$$(8) \quad \hat{\eta}_t^{EB} \approx \left(\frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{\sigma}_t^2} \right) \hat{\eta}_t ,$$

where $\hat{\eta}_t^{EB}$ is the EB estimate for teacher *t*, $\hat{\eta}_t$ is the initial estimate of effectiveness for teacher *t* based on the regression model (after combining OCCT estimates across grades), $\hat{\sigma}_t$ is the standard error of the estimate of teacher *t*, and $\hat{\sigma}$ is an estimate of the standard deviation of teacher effects (purged of sampling error), which is constant for all teachers. Equation (8) has no explicit term for the weight on the overall mean because this mean is zero. The term $[\hat{\sigma}^2/(\hat{\sigma}^2 + \hat{\sigma}_t^2)]$ must be less than one. Thus, the EB estimate always has a smaller absolute value than the initial estimate—that is, the EB estimate shrinks from the initial estimate. The greater the precision of the initial estimate—that is, the smaller $\hat{\sigma}_t^2$ is—the closer $[\hat{\sigma}^2/(\hat{\sigma}^2 + \hat{\sigma}_t^2)]$ is to one and the smaller the shrinkage in $\hat{\eta}_t$. Conversely, the larger the variance of the initial estimate, the greater the shrinkage in $\hat{\eta}_t$. By applying a greater degree of shrinkage to less precisely estimated teacher measures, the procedure reduces the likelihood that the estimate of effectiveness for a teacher falls at either extreme of the distribution by chance. We calculated the standard error for each $\hat{\eta}_t^{EB}$ using the formulas provided by Morris (1983). As a final step, we removed any teachers with fewer than 10 students and recentered the EB estimates on zero.

V. RESULTS FOR THE TLE SYSTEM

Educators will receive three types of results derived from the value-added results for teachers: (1) TLE component scores, (2) school value-added results, and (3) value-added results for student subgroups. In this chapter, we describe how we calculate each of these statistics from the value-added results.

A. Using the value-added results for the TLE system

We provided OSDE with the original generalized OCCT or EOI point value-added results, the average typical-peer and actual scores, and value-added results after converting them to a scale from 1.0 to 5.0 for each teacher in the model. The TLE system requires that the value-added results take on values from 1.0 to 5.0 when they are used to calculate composite TLE ratings. The state board approved the TLE Commission's recommended method of converting the value-added results to this TLE component scale. In this system, the average Oklahoma teacher received a score of 3.0; teachers whose results exceeded the average by two standard deviations received a score of 5.0; and those whose results fell below the average by two standard deviations received a score of 1.0. Teachers who were eligible for value-added results in multiple subjects received scores in each subject. We then assigned these teachers a combined component score that was a weighted average of their subject-specific scores. The weight given to each subject was the number of student equivalents.

B. School value-added results

Each school's value-added result reflects the combined contributions of teachers at that school. For each subject, we calculated a school's value added as a weighted average of the school's final teacher value-added results. The weight given to each teacher was the number of student equivalents. Rather than giving equal weight to each teacher in the calculation of school value added, this approach gives equal weight to students who have the same dosage in the value-added model. The average also includes value-added results for catch-all teachers of unassigned students.[8] As a final step, we calculated TLE component scores for schools across all subjects using the same procedure we used for teachers.

This method for calculating school value-added results provides two benefits compared to other frequently used methods. First, calculating school value added as a weighted average of teachers' value-added estimates is simpler and more transparent because it does not rely on estimating additional regression models, whereas other school value-added models typically do so. Second, the method addresses a source of potential bias related to how teachers are matched to students. Research suggests that disadvantaged students are often sorted to less effective teachers (Isenberg et al. 2013). This implies that schools with many disadvantaged students also have less effective teachers. Because they include school-fixed effects instead of the teacher-

[8] However, we excluded from the average the results for teachers who did not meet the 10 students reporting threshold for teacher value added. The alternative of including these results could lead to bias toward the average school's value added because the estimates for teachers below the reporting threshold receive a larger amount of shrinkage relative to those for teachers with more students.

fixed effects included in Equations (1) and (2), most other school value-added models do not account for the relationship between teacher effectiveness and student background.[9] Consequently, schools with many disadvantaged students receive school value-added results that might be too high. Aggregating teacher value-added results by school addresses this source of potential bias. Doing so accounts for the same correlation between the effectiveness of teachers and the characteristics of students in both the teacher and school value-added results.

C. Value-added results for student subgroups

We used the students' actual post-test scores and typical-peer scores to calculate the value-added results for student subgroups. The teacher's value-added result for a subgroup is the difference between the average actual post-test scores and the average typical-peer scores, where the averages are calculated based only on scores for students in the subgroup. The typical-peer scores reflect the adjustments that we made to the initial value-added estimates described in Chapter IV, including shrinkage and standardization across grades. We mean-centered the subgroup value-added results by combinations of subgroups and subjects so that a positive subgroup result reflects above-average contributions to the achievement of students in the subgroup compared to other Oklahoma teachers. As a final step, we removed subgroup value-added results for teachers or schools with fewer than seven students in the subgroup.

We produced value-added results for student subgroups that took on values of above average, average, or below average. We defined above average as value-added results that are greater than or equal to one standard deviation above the average teacher value-added result in the subject-subgroup combination. Similarly, we defined below average as less than or equal to one standard deviation below the average teacher value-added result in the subject-subgroup combination. We assigned a value of average to all other teachers with eligible subgroup value-added results. We produced results for student subgroups based on limited English proficiency status, individualized education program status, and proficiency levels on the pre-test from the same content area as the post-test.

[9] Because regressions with school-fixed effects use variation in achievement across teachers, these models can obtain coefficient estimates for the relationships between student characteristics and achievement that are larger in absolute value and lead to overadjusting for student characteristics. Guarino et al. (2012) discuss this possible consequence for the similar comparison between a teacher value-added model with teacher fixed effects to an alternative with no fixed effects.

REFERENCES

Arellano, Manuel. "Computing Robust Standard Errors for Within-Groups Estimators." *Oxford Bulletin of Economics and Statistics*, vol. 49, no. 4, November 1987, pp. 431–434.

Buonaccorsi, John P. *Measurement Error: Models, Methods, and Applications*. Boca Raton, FL: Chapman & Hall/CRC, 2010.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." Cambridge, MA: National Bureau of Economic Research, 2011.

Guarino, Cassandra Marie, Mark D. Reckase, and Jeffrey M. Wooldridge. "Can Value-Added Measures of Teacher Performance Be Trusted?" East Lansing, MI: Michigan State University Education Policy Center, December 2012.

Herrmann, Mariesa, Elias Walsh, Eric Isenberg, and Alex Resch. "Shrinkage of Value-Added Estimates and Characteristics of Students with Hard-to-Predict Achievement Levels." Washington, DC: Mathematica Policy Research, April 2013.

Hock, Heinrich, and Eric Isenberg. "Methods for Accounting for Co-Teaching in Value-Added Models." Washington, DC: Mathematica Policy Research, June 2012.

Isenberg, Eric, Jeffrey Max, Philip Gleason, Liz Potamides, Robert Santillano, Heinrich Hock, and Michael Hansen. "Access to Effective Teaching for Disadvantaged Students." Washington, DC: U.S. Department of Education, Institute of Education Sciences, November 2013.

Isenberg, Eric, and Elias Walsh. "Measuring Teacher Value Added in DC, 2012–2013 School Year." Washington, DC: Mathematica Policy Research, 2013.

Johnson, Matthew, Stephen Lipscomb, Brian Gill, Kevin Booker, and Julie Bruch. "Value-Added Models for the Pittsburgh Public Schools." Cambridge, MA: Mathematica Policy Research, February 2012.

Kane, Thomas J., and Douglas O. Staiger. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Working paper #14607. Cambridge, MA: National Bureau of Economic Research, 2008.

Kane, Thomas J., and Douglas O. Staiger. "Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains." Seattle, WA: Bill & Melinda Gates Foundation, 2012.

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment." Seattle, WA: Bill & Melinda Gates Foundation, 2013.

Liang, Kung-Yee, and Scott L. Zeger. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika,* vol. 73, no. 1, April 1986, pp. 13–22.

Morris, Carl N. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of American Statistical Association,* vol. 78, no. 381, 1983, pp. 47–55.

Pearson, Inc. *Oklahoma School Testing Program, Oklahoma Core Curriculum Tests, Grades 3 to 8 Assessments 2012 Technical Report.* Oklahoma City, OK: Pearson, Inc., 2012a.

Pearson, Inc. *Oklahoma School Testing Program, 2012 Technical Report, Achieving Classroom Excellence, End-of-Instruction Assessments.* Oklahoma City, OK: Pearson, Inc., 2012b.

Taylor, Eric S., and John H. Tyler. "The Effect of Evaluation on Performance: Evidence from Longitudinal Student Achievement Data of Mid-Career Teachers." Working paper #16877. Cambridge, MA: National Bureau of Economic Research, 2011.

Walsh, Elias, and Stephen Lipscomb. "Classroom Observations from Phase 2 of the Pennsylvania Teacher Evaluation Pilot: Assessing Internal Consistency, Score Variation, and Relationships with Value Added." Cambridge, MA: Mathematica Policy Research, May 2013.

www.mathematica-mpr.com

Improving public well-being by conducting high quality, objective research and surveys

PRINCETON, NJ ■ ANN ARBOR, MI ■ CAMBRIDGE, MA ■ CHICAGO, IL ■ OAKLAND, CA ■ WASHINGTON, DC

MATHEMATICA
Policy Research

Mathematica® is a registered trademark
of Mathematica Policy Research, Inc.