U.S. Department of Education
September 2013

Using alternative student growth measures for evaluating teacher performance: what the literature says

Brian Gill
Mathematica Policy Research

Julie Bruch
Mathematica Policy Research

Kevin Booker
Mathematica Policy Research

Key findings

States and school districts are exploring alternatives to state tests for measuring teachers' contributions to student learning. One approach applies statistical value-added methods to alternative student assessments such as commercially available tests and end-of-course tests. The evidence suggests that these methods can reliably distinguish among teachers. A second approach requires teachers to develop student learning objectives at the beginning of the school year; these can be used in instructional planning as well as evaluation. Ensuring consistency across teachers and schools is challenging, and implementation is demanding, but student learning objectives have the advantage that they can be implemented in any grade or subject.

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

REL MID-ATLANTIC
Regional Educational Laboratory
At ICF International

REL 2013–002

The National Center for Education Evaluation and Regional Assistance (NCEE) conducts unbiased large-scale evaluations of education programs and practices supported by federal funds; provides research-based technical assistance to educators and policymakers; and supports the synthesis and the widespread dissemination of the results of research and evaluation throughout the United States.

September 2013

This report was prepared for the Institute of Education Sciences (IES) under Contract ED-IES-12-C-0006 by Regional Educational Laboratory Mid-Atlantic administered by ICF International. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Gill, B., Bruch, J, & Booker, K. (2013). *Using alternative student growth measures for evaluating teacher performance: what the literature says.* (REL 2013–002). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from http://ies.ed.gov/ncee/edlabs.

This report is available on the Regional Educational Laboratory website at http://ies.ed.gov/ncee/edlabs.

Summary

States are increasingly interested in including measures of student achievement growth, or "value-added," in evaluating teachers. But annual state assessments, which are the typical measure of student growth, usually cover only reading and math teachers and only in grades 4–8. These state assessments thus cannot generally be used to measure contributions to student achievement growth for early elementary school teachers, most high school teachers, and teachers of other subjects.

As a consequence, a growing number of states and school districts are exploring alternatives for measuring teachers' contributions to student learning. These alternatives have the potential to be used for evaluating not only teachers who work in grades and subjects outside the annual state testing regime but also as complementary growth measures for teachers of tested grades and subjects.

This report reviews the literature on two categories of alternative measures for evaluating teachers:

- Alternative student outcome measures used in statistical growth (or value-added) models.
- Teacher-developed student learning objectives used for measuring growth.

Using alternative student outcome measures in statistical growth models

This literature review of studies of statistical growth models using alternative assessments (such as commercially available assessments like the Stanford Achievement Test and locally developed end-of-course exams) and other outcomes (such as student attendance) looked for evidence of the statistical properties of such measures. Despite differences in the student outcome measure, the statistical method used in the growth models to assess teacher value-added is similar to that used in state reading or math assessments. Key findings for growth/value-added models show that:

- Models based on widely used, commercially available assessments generally produce measures of teacher performance that correlate positively with other performance measures, such as teacher observations and student surveys. All the reviewed studies found positive relationships, with correlations up to 0.5.
- Models based on commercially available assessments yield results that are as stable over time as do models based on state assessments. Year-to-year correlations of teacher value-added based on commercially available assessments are positive but modest—consistent with year-to-year correlations for value-added measures based on state assessments. This finding suggests that growth models using these alternative measures—similar to those using state assessments—can be useful for teacher evaluation if applied judiciously. States and districts may want to use measures that average across several years of teaching or apply Bayesian "shrinkage" adjustments to reduce the likelihood that random error will mistakenly identify teachers as low performing.
- Little is known about growth/value-added models based on locally developed, curriculum-based assessments or nontest outcomes, but the available evidence suggests that they have the potential to reliably differentiate performance among teachers and schools. Just two studies were identified that examined the potential for using locally developed assessments to evaluate teacher performance, and both

examined the same district. The results suggest that such measures can reliably distinguish among teachers at the ends of the performance distribution. The same studies found that measuring schoolwide value-added using nontest outcomes (like attendance and course completion) can produce results that reliably distinguish school-level performance—but the studies did not analyze nontest outcomes at the teacher level.

More research is needed to inform the decisions of states and districts as they expand growth models to teachers and content not covered in state and commercially available assessments.

Measuring student growth using student learning objectives

Student learning objectives (SLOs)—classroom-specific growth targets chosen by individual teachers and approved by principals—are becoming popular as alternative measures of student growth because they can be used to evaluate teachers in any grade or subject. Although very little of the literature on SLOs addresses their statistical properties, key findings show that:

- SLOs have the potential to better distinguish teachers based on performance than traditional evaluation metrics do, but no studies have looked at SLO reliability. Most of the limited evidence on the statistical properties of SLOs is on the proportion of teachers achieving SLO objectives. Whether that differentiation represents true differences in teacher performance or random statistical noise is unknown.
- Little is known about whether SLOs can yield ratings that correlate with other measures of teacher performance. Only three studies have explored the relationship between SLO ratings and standardized assessment-based (value-added) growth measures. These studies found small but positive correlations. More research is needed as states and districts roll out SLOs as teacher evaluation measures and instructional planning tools.
- Until some of the research gaps are filled, districts that intend to use SLOs may want to roll them out for instructional planning before using them in high-stakes teacher evaluations. Several studies found teacher concerns about fairness in SLO implementation. This is no surprise, because SLOs are difficult to make valid and reliable. They are by definition customized to individual teachers and based on the professional judgments of teachers and principals. Making SLOs an important component of high-stakes evaluation could undermine their validity, because it means that teachers are in essence grading themselves.
- Studies of teacher experiences with SLOs indicate that SLOs can require substantial training and technology infrastructure and that they can be time-consuming for teachers and evaluators alike.

Contents

Summary i

Why this study? 1

What the study considered 1

Using alternative student outcome measures in statistical growth models 4
Criteria for selecting studies 5
Key findings on alternative outcomes used in statistical growth models 6
Gaps in the literature on alternative assessment–based value-added models 9

Measuring student growth using student learning objectives 10
Key findings on student learning objectives 11
Gaps in the literature on student learning objectives 13

Implications and further research 14

Appendix A. Literature search methodology A-1

Appendix B. Results of the literature search for alternative student outcomes in statistical growth models B-1

Appendix C. Results of the literature search for student learning objectives C-1

Notes Notes-1

References Ref-1

Tables

1 Number of correlations between alternative assessment and state assessment growth measures, by subject 7
2 Number of year-to-year correlations reported, by subject 8
A1 Search terms A-2
A2 Queries by alternative growth measure type A-2
A3 Database search results A-3
A4 Summary of citations identified A-3
A5 Summary of findings on implementation in districts and states A-4
B1 Key studies of alternative measures in growth models B-2
B2 Research on alternative student outcomes in growth models B-8
B3 Implementation of alternative student outcomes in growth models B-10
B4 All relevant studies identified on alternative measures in growth models B-11
C1 Statistical properties of student learning objectives C-2
C2 Implementation findings for student learning objectives C-4
C3 Implementation of student learning objectives C-7
C4 Data and methods in key student learning objective implementation studies C-8
C5 All relevant studies identified on student learning objectives C-9

Why this study?

Educators and policymakers nationwide are interested in measures of growth in student achievement, or teacher "value-added," as part of a system for assessing teacher and school effectiveness. All five jurisdictions in the Regional Educational Laboratory (REL) Mid-Atlantic Region have secured federal Race to the Top funding, which requires that their teacher evaluation systems incorporate measures of growth in student performance. The District of Columbia and Pennsylvania use value-added models (with multivariate regressions) to assess teachers' contributions to student learning, while Delaware, Maryland, and New Jersey rely on student growth percentiles. The methods differ somewhat, but both value-added models and student growth percentiles involve applying sophisticated statistical methods to data on the year-to-year changes in the achievement of individual students. As a matter of federal policy and for the purposes of this review, both are considered statistical growth models. This report uses student "growth models" and "value-added models" interchangeably.

The utility of growth/value-added models has been limited by the breadth of coverage and the depth of content in the underlying student assessments

The utility of growth/value-added models has been limited, however—and not just in the Mid-Atlantic Region—by the breadth of coverage and the depth of content in the underlying student assessments. Teacher-level growth models require student tests at least once a year, but most states have consecutive-year testing regimes only in grades 3–8 and only in reading and math. Relying solely on state assessments thus precludes measuring student growth for the majority of teachers serving other grades and subjects. Only about 15 percent of teachers in District of Columbia Public Schools, for example, can have their value-added measured using required assessments (TNTP, 2011). Even for reading and math teachers in tested grades, state assessments might not cover the whole curriculum and so might not capture the teachers' full contributions to student learning. Using state assessments for evaluation may give teachers incentives to focus their classes on topics covered in the state assessments.

The new state policy mandates have created an urgent need for additional measures of student achievement growth that can be applied in more subjects and grades and that can complement the content and format of state assessments (for example, by adding open-ended reading assessments that would complement existing multiple-choice vocabulary tests).

What the study considered

This research arose from interest expressed by REL Mid-Atlantic's Teacher Evaluation Research Alliance, which supports the development and refinement of teacher evaluation measures in the Mid-Atlantic states. The study reviews the research evidence on student growth measures other than those derived from statistical models applied to traditional state assessments in reading and math. The report summarizes the evidence on the use of three types of alternative growth measures:

- *Student assessments other than annual state assessments used in statistical growth models.* Some districts have used assessments that go beyond the standard statewide reading and math assessments in grades 3–8, including end-of-course exams (state-administered or locally developed) and widely used, commercially available assessments like the Stanford Achievement Test.
- *Nontest student outcomes used in growth models.* Some districts have measured school effectiveness using nontest outcomes, such as attendance, course completion, dropout, and graduation rates.

- *Student learning objectives (SLOs).* SLOs, adopted in at least seven states and six districts, are classroom-specific growth targets set by teachers and approved by principals. In this way they differ from growth models that rely on sophisticated statistical measures. For instance, a class might take an assessment at the beginning of the school year, and the teacher would set growth targets based on the students' performance and submit the targets to the principal for approval. The teacher would be evaluated on the proportion of students that achieve the target on an assessment given at the end of the school year. SLOs are becoming more popular in part because they can be used for measuring growth in any grade and subject.

The alternative student outcome measures in the first two categories are analyzed using value-added or student-growth percentile approaches, while the third category involves different ways of measuring a teacher's contribution to the outcomes as well as (often) different outcome measures. This report reviews the literature and summarizes the evidence on all three types of growth measures. It focuses on the statistical properties of the measures and the challenges in using them for evaluating teachers, two issues that are especially relevant for decisions by policymakers and educators on whether and how to use the measures.

There is a growing literature documenting year-to-year variability in teacher effectiveness measures based on standard value-added models

The literature review looked for three key statistical properties of alternative growth measures:

- How well they differentiate among teachers or schools.
- How reliable they are (the consistency of multiple scores on the same measure, or the absence of random measurement error).
- How they correlate with other measures of teacher effectiveness (such as value-added measures based on state assessments or measures of teachers' professional practice).

The ability of measures to differentiate among teachers is important because a key criticism of typical teacher evaluation measures is that they include only two rating categories: satisfactory and unsatisfactory—and nearly all teachers are rated satisfactory. Value-added models and student growth percentiles separate teachers into more categories or assign teachers scores that approximate a continuous distribution. The distribution of teachers across rating categories is thus the first characteristic to examine for alternative growth measures.

Evidence on reliability is necessary to determine that reported differences among teachers represent real differences rather than random statistical error. There is a growing literature documenting year-to-year variability in teacher effectiveness measures based on standard value-added models (McCaffrey, Sass, Lockwood, & Mihaly, 2009; Goldhaber & Hansen, 2010; Schochet & Chiang, 2010). Some random changes from year to year are inevitable, the result of the challenge of distinguishing a teacher's contribution from other factors correlated with student achievement (for example, students' prior schooling or classroom behavioral problems). Accounting for each student's prior achievement test scores means that a substantial percentage of what is left will be random error (or statistical "noise"). It is thus important to have enough students to produce a reasonably precise estimate of a teacher's contribution to student achievement growth. The report examines the limited evidence on the reliability of alternative growth measures.

Ideally, a literature review would uncover information on the validity of the alternative growth measures—in this case, how accurately a measure captures a teacher's performance rather than something outside the teacher's control, such as students' abilities. Validity is necessary for fairness. A measure lacking validity, such as one that gives some teachers higher ratings only because they happen to teach above-average students, would be unfair. Unfortunately, researchers rarely have a definitive benchmark of teacher performance as a test of validity—partly because observations of teachers' professional practice are imperfect and depend heavily on the observer, partly because student assessments cannot capture everything that students are expected to learn and teachers are expected to teach, and partly because students might be assigned to teachers in ways that are not fully accounted for by value-added models.

Absent a true measure of teacher effectiveness, this literature review includes evidence on how closely alternative growth measures correlate with other measures of teacher effectiveness

Absent a true measure of teacher effectiveness, this literature review includes evidence on how closely alternative growth measures correlate with other measures of teacher effectiveness. The bulk of the evidence compares value-added models estimated with alternative assessments against value-added models estimated with state assessments (for the same teachers).[1] In addition to reporting correlations between value-added models based on different assessments, the review also describes evidence on how closely value-added models estimated with alternative student outcome measures correlate with nontest measures, such as principals' ratings of teachers or student surveys on classroom climate and instruction.

Because the use of alternative growth measures is fairly new, the evidence base documenting their technical characteristics is small. Therefore, in addition to summarizing the literature, this report documents the gaps in knowledge about the measures' statistical properties. Some alternative outcome measures in use are underpinned by little or no research on their statistical properties. Documenting the gaps is useful both to inform states and districts about which measures have yet to be studied and to inform researchers seeking to expand their knowledge of the measures.

In addition to statistical properties, this study also reviews the literature on state and district implementation of the alternative measures. The review of implementation focuses primarily on SLOs. That is because the implementation challenges in applying statistical models to alternative outcome measures that school districts already use are similar to those in applying statistical models to state assessments: they require a good electronic data system and the technical capacity to conduct the analyses. By contrast, setting and measuring SLOs require time, training, and infrastructure beyond that required for value-added model analyses and place new demands on teachers and principals. Information on the benefits, drawbacks, challenges, and solutions, as well as teacher and principal attitudes toward SLOs, is useful as districts and states consider incorporating such measures into evaluation systems.

The literature search, including both qualitative and quantitative studies, aimed to be as inclusive as possible. Reviewed studies included discussion papers, news articles, publicly available academic papers, and reports produced by school districts and states. The systematic search included library databases, as well as the websites of states, districts, and other relevant institutions. (See appendix A for a detailed description of the literature search and results.) Identified studies were screened for inclusion based on relevance, evidence on alternative measures, and a focus on SLOs or a relevant statistical growth model. The

most relevant studies in each area were selected for review. All the studies and documents identified in the literature search are listed in appendixes B (statistical growth models) and C (SLOs).

The findings are presented first for the literature on applying statistical growth models to alternative student assessments and nontest student outcomes. The discussion of the two types of measures is combined because the methodological issues are similar and the empirical evidence on nontest outcomes is limited. The second section describes the literature on SLOs. Both sections describe the alternative growth measures in use, the statistical data on the measures, and the use of the measures for teacher evaluations.

Using alternative student outcome measures in statistical growth models

Widely used, commercially available assessments are a feasible alternative or complement to state assessments in value-added growth models

The reviewed studies indicate that widely used, commercially available assessments are a feasible alternative or complement to state assessments in value-added growth models. Models based on commercially available assessments generally yield measures of teacher performance that correlate with other performance measures and that are as reliable as models based on state assessments (sometimes even more reliable). Teachers who produce student growth on commercially available assessments also tend to produce growth on state assessments and tend to be ranked highly on other performance measures, such as classroom observations. All the reviewed studies found a positive relationship between measures of teacher performance.

The studies also indicate that teacher value-added based on commercially available assessments is about as stable over time as value-added based on state assessments. Nonetheless, as with value-added models using state assessments, models using alternative outcome measures include a substantial amount of random error that will cause results for some teachers to vary from year to year. This level of reliability results from random unmeasured differences between classes in student attributes like ability or background. Such random differences are accentuated when teachers have small classes and thus fewer test scores from which to draw inferences.

Averaging a teacher's value-added across years of teaching substantially reduces the variation in estimated performance. A Pittsburgh study found that 13 percent of middle school teachers could be distinguished from the average when value-added estimates were based on one year of data on the state reading assessment; that share rose to 36 percent when estimates were based on three years of data (Lipscomb, Gill, Booker, & Johnson, 2010). Another way to reduce the likelihood of mistakenly identifying teachers as low performing due to random error is to use a Bayesian shrinkage adjustment, which implicitly assumes that a teacher is average unless strong evidence shows otherwise. Standard value-added models usually apply a shrinkage adjustment, which can be used in combination with averaging performance across multiple years to increase the stability of results.

These findings on the validity and reliability of value-added estimates for commercially available assessments show the potential for including them in teacher evaluation measures. Because of the high correlation between value-added estimates using different types of assessments, a teacher evaluation that incorporates value-added results from both commercially available assessments and state assessments should be more stable over time than an evaluation that incorporates just one type of assessment. Similarly, because of

the correlation with other performance measures, value-added estimates for widely used, commercially available assessments could make a composite performance measure incorporating those other performance measures more stable than one that used only value-added estimates for state assessments. Empirical verification of the stability of composite measures was a key finding of the newly released final reports of the Bill & Melinda Gates Foundation's Measures of Effective Teaching project (Bill & Melinda Gates Foundation, 2013).

Much less is known about measuring growth based on locally developed, curriculum-based assessments or on nontest student outcomes. Only two studies, by the same research team and examining data for Pittsburgh, have examined the use of value-added models incorporating locally developed curriculum-based assessments. Both found encouraging results in the ability to reliably distinguish among teachers at the ends of the performance distribution (Johnson, Lipscomb, Gill, Booker, & Bruch, 2012; Lipscomb et al., 2010). The same studies were also the only ones to have examined the application of value-added models to nontest student outcomes, such as attendance and course completion. The studies again found evidence that value-added models using those measures can produce results that reliably distinguish performance—but the studies explored these outcomes in value-added models at the school level, not at the teacher level.

Most of the evidence compares teacher value-added on state assessments with value-added on commercially available assessments

Detailed findings are discussed below. Appendix B provides additional information on the studies.

Criteria for selecting studies

The analysis was based on the 14 studies on alternative growth models identified in the literature search that had the most information on statistical properties of measures and that were the most relevant for educators and policymakers (see table B1 in appendix B). In addition to meeting the search criteria (appendix A), the 14 studies met the following additional criteria:

- Their statistical models of student growth/value-added rely on longitudinal data on individual students, thereby controlling for students' prior achievement.
- With one exception (discussed below), they estimate value-added at the teacher level.

The analysis determined whether a study was subjected to a formal or informal peer review (to the extent that could be ascertained) and distinguished formal peer reviews by academic journals from those by government agencies, internal organizations, dissertation committees, and other researchers (see table B1).

Most of the evidence compares teacher value-added on state assessments with value-added on commercially available assessments, such as the Stanford Achievement Test, the Balanced Assessment in Mathematics, and the Scholastic Reading Inventory. Most of the studies focus on the Stanford Achievement Test, though different versions are used in different locations. The districts participating in the Bill & Melinda Gates Foundation's Measures of Effective Teaching project used an open-ended reading assessment, while other locations used a multiple-choice version. Two studies compare teacher value-added on traditional state assessments with value-added on end-of-course exams. One study explores value-added on PSAT scores, and one looks at nontest outcomes such as attendance and

core course pass rates but with schools (rather than teachers) as the unit of analysis. The study is nonetheless included in the analysis because it is the only one that uses nontest outcomes in value-added models.

Key findings on alternative outcomes used in statistical growth models

This section presents detailed information on the correlation of student growth/value-added models using alternative outcomes with other performance measures, on the reliability of the alternative measures, and on other key findings.

Findings on correlation with other measures. Growth/value-added models based on commercially available assessments generally yield measures of teacher performance that correlate with other performance measures.

Value-added models based on commercially available assessments generally yield measures of teacher performance that correlate with other performance measures

Teacher value-added using commercially available low-stakes assessments correlates moderately with value-added from state assessments. Several studies examined the correlation between value-added on traditional state assessments and value-added on alternative assessments (for the same teachers). A correlation of 0 indicates no relationship between the two measures of the same teacher, and a correlation of 1 indicates that the two measures produce identical results for each teacher. Four studies of teacher value-added in grades 4–8 published by the Bill & Melinda Gates Foundation examined correlations between two alternative tests (the Balanced Assessment in Mathematics and the Stanford-9 Open-Ended Reading) and the state tests (Bill & Melinda Gates Foundation, 2010, 2012; Kane et al., 2013; Mihaly, McCaffrey, Staiger, & Lockwood, 2013). The overall correlation for teacher value-added between the alternative assessment and the state assessment ranged from 0.41 to 0.54 in math and from 0.39 to 0.59 in reading. A similar comparison of teacher value-added for grades 4–5 in Houston for the Stanford Achievement Test and the state assessment found correlations of 0.59 for math and 0.50 for reading (Corcoran, Jennings, & Beveridge, 2011).

A comparison of the Scholastic Reading Inventory and state assessments for grade 3–5 reading in a large Northeastern school district found smaller correlations in value-added estimates of 0.16–0.44 (Papay, 2011). The lower correlations might have been due to comparing student gains from fall to fall with those from spring to spring. And a comparison of the Scholastic Reading Inventory and the state reading assessments in Pittsburgh found a correlation in value-added estimates of 0.25 (Lipscomb et al., 2010).

Most of the correlations in the 14 studies involve teacher-level value-added estimates, but a few involve school-level estimates. Each correlation was positive, though most were not large, especially in reading, where 9 of 11 correlations were less than 0.5 (table 1).

Teacher value-added using commercially available low-stakes assessments correlates with measures of teaching practice, principals' judgments, and student surveys, and the correlation is about as high as for state assessments. A study of more than 1,000 teachers in six districts compared the correlation for a classroom observation measure and teacher value-added estimated through both low-stakes alternative assessments (the Balanced Assessment in Mathematics and the Stanford 9 Open-Ended Reading) and the state assessment. It found a stronger correlation with classroom observations for the alternative assessments than for the state assessment (Bill & Melinda Gates Foundation, 2012). A follow-up study in the

Table 1. Number of correlations between alternative assessment and state assessment growth measures, by subject

Alternative assessment	Math			Reading		
	Negative	Positive, < 0.5	Positive, > 0.5	Negative	Positive, < 0.5	Positive, > 0.5
All	0	5	2	0	9	2
Stanford[a]	0	5	2	0	6	2
Scholastic Reading Inventory	na	na	na	0	2	0

na is not applicable.

Note: Some of the studies did not report the statistical significance of the correlations, so the coefficient estimates are counted instead.

a. One study pooled its estimates for math and reading; the correlation is included in both the math and reading columns.

Source: Authors' analysis of 14 studies listed in table B1 in appendix B.

The reliability of teacher value-added using commercially available assessments is comparable to that using state assessments

same districts found a similar pattern, with the exception of elementary school math, for which value-added based on the state assessment had a stronger correlation with a teacher observation measure (Mihaly et al., 2013).

The Bill & Melinda Gates Foundation (2012) study also found that teacher value-added as measured by the widely used, commercially available alternative assessments correlated with a composite measure of teacher performance that included value-added on the state assessment, measures of observational practice, and results of student surveys.

Another study, examining the correlation between value-added based on the Stanford Achievement Test and teacher ratings based on interviews with principals about teacher characteristics, found moderate correlations (0.15–0.34) in math and reading (Harris & Sass, 2012). "Motivation/enthusiasm" had the strongest correlation with reading value-added, while "knowledge/teaching skills/intelligence" had the strongest correlation with math value-added. The study found more overlap between value-added estimates and principal ratings at the bottom end of the distribution than at the top: 65 percent of teachers that ranked in the bottom 30 percent based on value-added on the math Stanford Achievement Test were also ranked in the bottom 30 percent by principals, while only 16 percent of those ranked in the top 30 percent based on math value-added were also ranked in the top 30 percent by principals.

Findings on reliability of alternative measures. Most of the studies also examined the reliability of value-added estimated with alternative outcome measures.

The reliability of teacher value-added using commercially available assessments is comparable to that using state assessments. Models based on commercially available assessments generally yield measures of teacher performance that are as reliable as value-added models based on state assessments. Several studies examined reliability by assessing the agreement of growth measures for the same teacher across years or student groups. A Florida study compared year-to-year correlations in teacher value-added in grades 4–10 for five school districts on the Stanford Achievement Test and the state assessment (McCaffrey et al., 2009). Florida mandated the Stanford Achievement Test as a commercially available assessment to complement the state assessment. The study found that year-to-year correlations varied from 0.2 to 0.5 for elementary school teachers and from 0.3 to 0.6 for middle school teachers,

with no clear pattern of differences between year-to-year correlations on the alternative assessment and the state assessment.

A Texas study comparing the persistence of teacher value-added on the Stanford Achievement Test and on the state assessment found the Stanford value-added to be more persistent over time (Corcoran et al., 2011). About 60 percent of the teacher's value-added on the Stanford carried over from one year to the next, compared with about 40 percent on the state assessment. Comparisons of value-added that were two years apart showed a similar pattern, with 53 percent carryover on Stanford value-added and 32 percent on the state assessment.

Across studies, year-to-year correlations of value-added estimates based on alternative assessments were almost entirely positive (except for one local reading assessment; table 2). The correlations are often consistent with those for state assessments (for example, McCaffrey et al., 2009) and with estimates of year-to-year correlations of measured performance for complex professions outside teaching (such as managers and research scientists), which typically range from 0.33 to 0.40 (Sturman, Cheramie, & Cashen, 2005, as cited in Glazerman et al., 2010). But some studies found more modest year-to-year correlations for teachers, at less than 0.3, suggesting somewhat greater random variation than is typically found in other fields.

Across studies, year-to-year correlations of value-added estimates based on alternative assessments were almost entirely positive

A Pittsburgh study found that value-added estimates based on local assessments designed to align with the curriculum (though not developed by psychometric experts) are reliable enough to reveal statistically significant distinctions among teachers (Johnson et al., 2012). Instead of measuring year-to-year reliability, the study used student-level variation in predicted and actual achievement for all of a teacher's students. This gets at the extent to which measures for the same teacher are consistent across students. If there is substantial noise across students, it would be difficult to distinguish teachers from the average. Using data from a variety of locally developed assessments in Pittsburgh Public Schools, the study distinguished 36 percent of teachers from the district average (at a 95 percent confidence interval). In comparison, the study distinguished 28 percent of teachers from the district average using state assessments.

Table 2. Number of year-to-year correlations reported, by subject

	Math			Reading			Other measures[a]		
Alternative assessment	Negative	Positive, < 0.3	Positive, > 0.3	Negative	Positive, < 0.3	Positive, > 0.3	< 0	0–0.3	> 0.3
All	0	5	2	1	2	4	0	0	3
Stanford[b]	0	4	1	0	2	3	—	—	—

— is not available.

Note: Some of the studies did not report the statistical significance of the correlations, so the coefficient estimates are counted instead.

a. Includes science, social studies, and writing alternative assessments, as well as attendance and credits earned.

b. One of the studies used the Stanford 9 Open-Ended Reading Assessment; the others used multiple-choice versions of the Stanford Achievement Test.

Source: Authors' analysis of 14 studies listed in table B1 in appendix B.

Other key findings on alternative outcomes used in statistical growth models

Other study results can also inform decisions by policymakers and educators on whether and how to use alternative outcomes in growth/value-added models.

Teacher value-added using end-of-course exams in secondary school grades could be sensitive to student tracking by ability. A North Carolina study that looked at teacher value-added for grade 9 using end-of-course exams in algebra and English found substantial student tracking into "high ability" and "low ability" groups (Jackson, 2012). If not taken into account, this kind of tracking could undermine the validity of value-added estimates. Including statistical controls for student tracking (as in Johnson et al., 2012) may be important for avoiding bias in value-added estimates in secondary school.

Little is known about the alternative measures' reliability or correlations with other measures in other locations using alternative growth measures

Little evidence exists on the application of statistical growth models to nontest outcomes. Only one study applied value-added models to nontest student outcomes (Johnson et al., 2012). The Pittsburgh study looked at attendance for grades 1–12 and core course pass rates for grades 9–12 to estimate schoolwide value-added but not teacher value-added. Some schools could be reliably distinguished from others in value-added for attendance. The distribution of school effects based on attendance increased with grade level, from a 0.05 *z*-score gap between the mean and 90th percentile in grades 1–3 to a 0.21 gap in grades 9–12. This suggests that using student attendance as an outcome in value-added models is more useful in high school than in lower grades. Note, however, that there is no evidence on the use of attendance or other nontest outcomes in teacher (as opposed to school) value-added models.

Gaps in the literature on alternative assessment–based value-added models

Although the review uncovered 14 studies that examined the statistical properties of alternative assessment measures (of the 44 relevant studies identified; see table B4 in appendix B), large gaps remain. The literature is dominated by research on the use of the Stanford Achievement Test in value-added models; less is known about other assessments. Nine of the fourteen studies report on the statistical properties of Stanford-based value-added models (see table B1). Two studies also examine the Balanced Assessment in Mathematics, two examine the Scholastic Reading Inventory, one examines the Measures of Academic Progress, and one examines Dynamic Indicators of Basic Early Literacy Skills and Terra Nova. Four studies examine various secondary-level end-of-course exams, including locally developed assessments.

The evidence base is also limited geographically, with Florida overrepresented, perhaps because it has had a high-quality statewide student data system for some time. Half the 14 key studies include at least one Florida district. Three studies focus primarily on Florida districts; the other four include districts in the Bill & Melinda Gates Foundation's Measures of Effective Teaching project, which includes Florida's Hillsborough County. The studies presented evidence on the statistical properties of measures used in other locations, such as the distribution of teacher effects (see table B4 in appendix B), but little is known about the alternative measures' reliability or correlations with other measures in other locations using alternative growth measures (including Ohio, Tennessee, Charlotte-Mecklenburg, Detroit, Milwaukee, Little Rock, Tulsa, the EPIC charter school consortium, and the SIATech charter school network; see table B2). Some of these locations began using

alternative growth measures in 2011/12, and some in 2012/13. Others have used alternative growth measures for several years. More data are being produced in these locations, which will allow for further research.

The bulk of this review focuses on alternative assessments and SLOs, but three locations also use nontest measures, such as attendance and core course pass rate, for evaluating teachers or schools (see table B1). Some evidence was found on the statistical properties of these measures in Pittsburgh but not in other locations.

Measuring student growth using student learning objectives

SLOs are another way to measure student growth. As classroom-specific achievement targets, SLOs are typically set by individual teachers, usually in consultation with their principals, and may be based on any of a wide range of student assessments, including state assessments, commercially available assessments, and teacher-developed (nonstandardized) assessments. These measures are distinct from the growth models described in the previous section because they do not rely on sophisticated statistical methods for attributing student achievement growth to teachers. Rather, they are based on teachers' and principals' knowledge of individual students and assumptions about students' expected growth during a school year. SLOs are becoming more popular with states and districts looking for growth measures to include in newly mandated teacher evaluation metrics—especially in grades and subjects not covered by state or commercially available assessments.

Most of the evidence on student learning objectives consists of lessons from implementation; there is little evidence on their statistical properties

Only seven studies on SLOs included data on reliability, validity, or the percentage of teachers meeting SLOs (see table C1 in appendix C, which also lists the locations). Seven studies included implementation lessons based on data collected from teachers or districts (see table C2). And three of these seven are among the five with data on statistical measures; four other studies focused only on implementation. Reports that did not document the systematic collection of empirical data were excluded. The studies in table C2 provide information on teacher attitudes toward SLOs, teachers' opinions of the relevance and usefulness of SLOs, principals' attitudes toward the workload involved, and the challenges encountered in terms of assessments and time. (See appendix C also for more information on where SLOs are being used, the data and methods used in each key implementation study, and a list of all the studies and documents identified in the literature search.)

Most of the evidence on SLOs consists of lessons learned from implementation in various locations; there is little evidence on the statistical properties of SLOs, particularly on correlations with value-added measures and on year-to-year reliability. The evidence identifies some key areas for implementing SLOs, including ensuring that teachers have proper training and appropriate tools for creating SLOs and tracking data and taking into account validity concerns likely to arise when teachers set SLO targets. These implementation challenges might prompt districts and states to roll out SLOs in a low-stakes context before including them in high-stakes teacher evaluation systems.

Most of the districts implementing SLOs are using them for more than evaluation. Many teachers value SLOs for professional development and planning, which suggests there are advantages to introducing SLOs for these purposes before using them for (or alongside) high-stakes evaluations. Districts planning to use SLOs for instructional improvement should be aware that attaching consequences to the achievement of SLOs risks producing

pressure for "grade inflation," potentially undermining the value of the SLOs for instructional planning.

More generally, districts and states should understand that SLOs are difficult to make valid and reliable. They are by definition customized to individual teachers and based on the professional judgments of teachers and principals—who typically have neither the data nor the training to promote consistency and rigor in their use. The fact that teachers have a stake in the results is likely to exacerbate the difficulties, because it means that teachers are in essence grading themselves. States and districts might begin to tackle these challenges by ensuring that principals have a strong role in setting SLO targets. Principals have the advantage of seeing many teachers across the school and have less of an incentive to make the SLOs easy to reach. This does not, however, address the problem of consistency across schools. Districts should consider some sort of centralized auditing to assess and promote a consistent level of rigor in SLO targets across schools and teachers.

Student learning objectives are difficult to make valid and reliable

Key findings on student learning objectives

This section summarizes key findings on statistical properties and implementation of SLOs and identifies the main gaps in the literature.

There is little evidence on the statistical properties of SLOs. Most of the limited evidence on the statistical properties of SLOs simply reports the proportion of teachers achieving SLO objectives (see table C1). Across various sites the results consistently show that most teachers achieve some or all of their SLO targets. In Denver, early in the implementation of the district's professional development and compensation program, ProComp, 89–98 percent of participating teachers met at least one SLO (Community Training and Assistance Center, 2004). A few years later, a large majority of teachers continued to meet SLOs, with 70–85 percent of participating teachers earning a financial incentive tied to meeting them (Goldhaber & Walch, 2011; Proctor, Walters, Reichardt, Goldhaber, & Walch, 2011). Studies in Tennessee (Tennessee Department of Education, 2012) and Austin, Texas (Terry, 2008), found that about two-thirds of teachers met all of their SLO targets (involving one goal in Tennessee and two in Austin). In Charlotte-Mecklenburg attainment rates ranged from 55 percent to 81 percent over three years (Community Training and Assistance Center, 2013). While in Charlotte-Mecklenburg teachers' success rates rose in the second year of implementation and fell in the third year (Community Training and Assistance Center, 2013), success rates in Denver gradually increased over 2007–10 (Goldhaber & Walch, 2011).

These results suggest that SLOs may better discriminate among teachers than do traditional evaluation metrics (in which nearly all teachers are deemed "satisfactory"). Still, more than half the teachers met their targets in all the locations studied. And considering that the two Community Training and Assistance Center studies (2004, 2013) found that success rates increase the longer teachers participate in an SLO program, districts could find a large majority of teachers rated in the highest category several years after SLO implementation begins.

Whether the ability of SLOs to distinguish among teachers represents true differences in teacher performance or random statistical noise remains to be determined. No studies have attempted to measure the reliability of SLO ratings.

Although there are no measures of the reliability of SLO ratings, three studies have begun to explore the validity of SLO ratings by examining their relationship with standardized test–based measures. A Denver study found a small but statistically significant positive relationship between the receipt of ProComp awards and teacher value-added (Goldhaber & Walch, 2011). Slightly more teachers in the top two quintiles of value-added performance earned SLO-based awards than did teachers in the bottom two quintiles (42 percent to 37 percent in math and 41 percent to 38 percent in reading). An analysis in Austin of the relationship between meeting objectives and net growth on the Texas state assessment found that teachers that met at least one objective outperformed those that met none (Schmitt & Ibanez, n.d.). This was true overall and for novice teachers. A study that estimated the relationship between meeting SLOs and student achievement on end-of-grade exams, controlling for students' prior achievement and other characteristics, also found a positive relationship (Community Training and Assistance Center, 2013).

Teachers need guidance and appropriate tools to implement student learning objectives successfully

Teachers need guidance and appropriate tools to implement SLOs successfully. Based on teacher surveys and focus groups, several studies found that teachers reported needing support in setting and implementing SLOs. An Austin Independent School District (2012) study found that teachers requested additional guidance on the SLO assessment process and that some participants were unfamiliar with the measures in use. A study of SLOs in Denver found that teachers initially considered the SLO-setting process to be complex and needed greater support and feedback (Community Training and Assistance Center, 2004).

An examination of an SLO pilot in Austin found that technology problems led to frustration and the perception among teachers that the system was faulty and difficult to use (Burns, Gardner, & Meeuwsen, 2009). After analyzing teacher survey and interview data, the Community Training and Assistance Center (2004) recommended that districts rolling out SLOs focus on linking student information systems and human resources information systems to make the SLO data more useful in teacher evaluations. Based on interviews with district and state officials, the Reform Support Network (2012) emphasized teachers' need for tools to help them develop SLOs, particularly in setting rigorous goals, and for improved systems to limit data errors.

The Tennessee Department of Education (2012) reported that some teachers did not find the SLOs consistent with their job responsibilities. TNTP (2012) and Community Training and Assistance Center (2013) emphasized the importance of data accessibility to the quality of SLOs. TNTP found that getting prior-year student data to create starting points was a major challenge and recommended storing such data in a way that gives teachers easy access. The study also emphasized identifying and setting SLOs before the beginning of the school year. The Community Training and Assistance Center study also identified technology as a key component, noting that one of the main distractions identified by teachers was an inadequate and constantly changing software platform for storing SLO data.

SLOs require more time of teachers and evaluators. Teachers participating in Indiana's RISE evaluation system, which includes SLOs, spent more time measuring student learning (a median of 6.5 hours) than did non-RISE teachers (4 hours; TNTP, 2012). RISE participants noted that creating and updating assessments for SLOs took longer than any other part of the process. More than a third of teachers who created their own assessments spent at least 5 hours on it. Developing assessments also requires evaluators' time: RISE evaluators spent

a median of 30 minutes per teacher providing feedback on the teacher-developed SLO assessments. In Denver principals were split, with 28 percent reporting that SLOs raised their administrative workload and 34 percent reporting the opposite.

SLOs might have value for instructional improvement apart from their use for teacher evaluation. As well as a teacher evaluation measure, SLOs can be a tool for instructional improvement, with the aim of focusing teachers' attention on students' needs and goals. A study of SLO implementation in Denver found that more than 60 percent of teachers believed that SLOs improved their instructional practices, and close to 60 percent believed that SLOs will raise student achievement (Proctor et al., 2011). Similarly, an Austin study found that 67 percent of teachers agreed that SLOs are a positive change, with some teachers reporting that SLOs helped them target their goals and identify areas in need of improvement (Schmitt & Ibanez, n.d.). A Charlotte-Mecklenburg study found that teachers and principals used SLOs to improve instruction (Community Training and Assistance Center, 2013). By contrast, Tennessee Department of Education (2012) found that teachers did not view teacher-selected goals as drivers of effective instruction. The Austin study also found that some teachers reported spending large amounts of time on SLOs but that they did not always see a clear link between doing so and instructional improvement.

Much more evidence is needed to understand the reliability and validity of student learning objectives for evaluation

The inherent inconsistency of SLOs across teachers raises concerns about the validity of SLOs as teacher evaluation measures. Several studies found teacher concerns about fairness in SLO implementation. Austin Independent School District (2012) reported that some focus group participants were frustrated that student mobility, dropout, and attendance had different impacts on teachers' ability to meet SLO goals. Tennessee Department of Education (2012) found that teachers viewed the SLOs as the least effective component of the evaluation system, in part because similar groups of teachers did not consistently select the same measures. The study found that assessment choices were often based on the teachers' and principals' beliefs about which assessments would produce the highest scores. A Denver study likewise identified teacher concerns about the consistency of implementation of SLOs across schools (Proctor et al., 2011). In Austin, where two-thirds of teachers viewed SLOs favorably for instructional purposes, two-thirds also disagreed that SLOs are a good measure of effective teaching (Burns et al., 2009). Unsurprisingly, Austin teachers who met their SLO targets were far more likely to consider SLOs to be good measures of effective teaching.

Gaps in the literature on student learning objectives

The dearth of evidence on the statistical properties of SLOs is the most notable gap in the literature. Several studies included information on the percentage of teachers meeting SLOs, but they covered only four locations (Austin, Charlotte-Mecklenburg, Denver, and Tennessee), and none examined the reliability of SLO ratings. Only three studies examined the correlation of SLOs with other measures of teacher effectiveness (Goldhaber & Walch, 2011, in Denver; Schmitt & Ibanez, n.d., in Austin; and Community Training and Assistance Center, 2013, in Charlotte-Mecklenburg). These studies found evidence of modest positive relationships between meeting SLO targets and student achievement on state assessments. Much more evidence is needed to understand the reliability and validity of SLOs for evaluation.

Another area lacking information is the variability of SLO standards across teachers and schools, indicated by findings in Denver (Proctor et al., 2011) and Tennessee (Tennessee Department of Education, 2012). Teacher concerns about the consistency of standards for setting SLOs seem appropriate, given that targets are determined by the individual judgments of teachers and principals. To the extent that the achievement of SLO targets is used for high-stakes decisions (such as financial bonuses), teachers will have strong incentives to lower expectations and make their SLO targets easier to reach. The issue merits more attention by researchers and highlights the close relationship between the statistical properties of SLOs and implementation issues.

The geographic coverage of the key studies on SLO implementation (see table C2) is limited, with half the studies focusing on Denver or Austin.

The evidence base on the use of alternative student growth measures remains small

The limited evidence on SLO implementation also raises questions. It is possible that apparently contradictory views of SLOs can be explained by the two different purposes of SLOs: instructional planning and teacher evaluation. Whether both purposes can be served at once is a key question. In a high-stakes teacher evaluation context the pressure to make SLO targets easier to reach could undermine their value for instructional planning and improvement.

As more SLOs are implemented, more opportunities to learn from them will arise. In particular, as various locations implement different SLOs with different types of guidance and training, more information will become available about what works and how to approach common challenges.

Implications and further research

Despite increasing steadily, the evidence base on the use of alternative student growth measures remains small. Studies that include evidence on the statistical properties of the measures are limited almost exclusively to cases where standard value-added methods are applied to tests other than the state's high-stakes assessments. Most often, the studies use commercially available alternative assessments, such as the Stanford Achievement Test, which are rarely used for high-stakes teacher accountability.

Assessing teachers based on student growth on these assessments typically yields results that distinguish teachers from each other, that are comparably reliable to value-added estimates based on state assessments, and that correlate positively both with value-added on state assessments and with other performance measures, such as classroom observations. Districts and states looking for additional student growth measures could consider applying value-added statistical methods to commercially available assessments that align with their curriculum and standards.

Even less evidence is available for alternative growth measures that apply value-added methods to other student outcome measures, such as locally developed end-of-course assessments. The limited evidence available (mostly from Pittsburgh) is promising for teacher differentiation, reliability, and correlations with other measures. While locally developed assessments are unlikely to be as reliable as widely used, commercially available assessments, they have the potential to align more closely with the local curriculum and standards.

Any single performance measure based on student growth in a single year will be statistically noisy, and ratings based on such measures will fluctuate from year to year. The annual fluctuation can be mitigated by incorporating multiple years of performance into a teacher's assessment, by combining value-added results from alternative assessments and other performance measures, and by applying a Bayesian shrinkage adjustment to the growth measure (see, for example, Bill & Melinda Gates Foundation, 2013; Mihaly et al., 2013).

States and districts considering incorporating alternative student growth measures into their teacher evaluation systems should find it encouraging that these measures are fairly reliable (especially if averaged across several years of teaching) and positively correlated with other measures. The evidence is also positive for growth models using end-of-course exams if student ability tracking is accounted for. Applying value-added methods to end-of-course assessments will be especially valuable at the high school level, where traditional reading and math accountability tests struggle to account for the diversity of course content and the range of grade levels within classrooms.

Many districts and states cannot wait until the research base fills out before implementing some kind of alternative growth measure, particularly for teachers in grades and subjects not covered in state assessments

For SLOs, most of the evidence is on implementation, with little on statistical properties. The fact that SLOs are devised by individual teachers and principals suggests that both reliability and correlation with other measures would be tough expectations to meet—and explains why teachers often doubt the measures' fairness for evaluation. Nonetheless, the limited statistical evidence points to positive relationships between achievement of SLO goals and student achievement on state assessments. Evidence on the reliability of SLO ratings has yet to be produced.

Studies of SLO implementation make clear that it can be a demanding process. Teacher and principal training is likely needed to instill rigor in SLO goals. Districts and states should anticipate that implementing SLOs will create more work for teachers and principals. Data systems need to be accessible and responsive so that teachers can understand their students' starting points and set realistic growth targets. And the fairness and consistency of implementation across teachers and schools often become matters of concern. Researchers have much work to do to learn how these challenges play out in different contexts and how to overcome them.

Many districts and states cannot wait until the research base fills out before implementing some kind of alternative growth measure, particularly for teachers in grades and subjects not covered in state assessments. The evidence on the application of value-added methods to alternative student tests is encouraging, suggesting that states and districts might want to begin by seeking out commercially available assessments that align with their curricula and standards or by seeking (or even developing) end-of-course assessments that are centrally scored and align with curricula and standards. The same value-added statistical methods used on state assessments can be used on other kinds of systematically scored student tests.

Districtwide or statewide alternative assessments, however, are unlikely to work in all grades and subjects. SLOs have become more popular partly because they provide a way to address gaps in the testing regime, partly because their individualized design can grant teachers considerable autonomy, and partly because some teachers find them useful for instructional planning and improvement. But there is likely to be tension between using SLOs for instructional improvement and using them for high-stakes evaluation. Teachers

themselves are setting the targets, and if their evaluation depends on reaching the targets, teachers will have an incentive to set the targets low. So districts need to recognize that the efficacy of SLOs for instructional improvement could be undermined in high-stakes contexts.

More generally, because teachers can customize SLOs, it is difficult to use them fairly for evaluation. Their validity as measures of teacher performance depends on reasonable consistency in how difficult they are to achieve. Wide variation in the rigor of SLO targets among teachers within or across schools could vitiate their usefulness for teacher evaluations and be unfair to teachers who set high expectations.

No evidence exists on how to solve this problem entirely. But standardizing the process within and across schools might be one way to mitigate the adverse effects. Other potential ways to improve consistency of difficulty within and across schools is to authorize principals to modify SLO targets and to institute district training systems and auditing systems. Studies are needed to determine whether such measures would be sufficient to ensure the validity of SLOs for teacher evaluation. Without such systems it is likely to be nearly impossible to make valid and reliable comparisons of teachers using SLOs.

Districts need to recognize that the efficacy of student learning objectives for instructional improvement could be undermined in high-stakes contexts

Appendix A. Literature search methodology

This literature review set out to describe the landscape of alternative measures of student achievement used in growth models and to document studies reporting empirical evidence on statistical properties. Because alternative student outcome measures have only recently been included in growth models, few empirical studies exist. The search aimed to be as inclusive as possible.

Types of studies identified

The search covered both qualitative and quantitative studies. The qualitative data, drawn from discussion papers, news articles, and other documents, were used to identify measures in use for which there is not yet quantitative empirical evidence or to describe the implementation, including the logistical advantages and disadvantages of each measure. The quantitative data were drawn from publicly available academic papers and reports produced by school districts and states.

The study began with a systematic search of the library databases EBSCO Education Research Complete and EconLit. Because much of the literature comprised reports not published in journals, the search was expanded to include the more inclusive Google Scholar search engine. Also searched were the websites of states and districts that have used or are planning to use alternative student outcomes in growth models. Citations in the reviewed studies identified other states and districts implementing these measures, and they were added to the search. Searches were conducted of the websites of the Council of Chief State School Officers, the Education Commission of the States, the National Comprehensive Center for Teacher Quality, the Teacher Incentive Fund Community, and the Value-Added Research Center as well.

This topic is fairly new, so terminology varies across sources, necessitating a wide range of search keywords. Five categories of search terms were identified and combined in Boolean queries to target the most relevant citations (table A1).

To limit the number of results outside the scope of the review, queries were defined so that articles must contain either an alternative assessment term or a nontest measure term within 10 words of at least one student growth term (table A2). Because the nontest measure terms are less specific, articles containing them also had to include at least one evaluation or performance term anywhere in the full text.[2] Queries for student learning objectives (SLOs) did not include a student growth term, but they did include an evaluation or performance term.

Potentially eligible studies were screened for inclusion based on the following criteria:

- Discussed a measure of student achievement growth based on an assessment or a nontest outcome (attendance, course completion, dropout, and graduation).
- Did not focus solely on state standardized tests in grades 3–8 reading and math as the outcome variable for students.
- Focused on SLOs or employed a growth model intended to isolate teachers' or schools' contribution to student growth. For a study to be included, its growth models had to be used to measure teacher or school effectiveness and had to rely on longitudinal student data. Measures that rely on changes in aggregate

Table A1. Search terms

Category	Search term	
Student growth	• Value-added • Student growth	• Residual gain
Alternative assessments	• End of course • End of semester • End of year • Curriculum-based assessment • Curriculum-based test • Iowa Test of Basic Skills/ITBS	• Stanford Achievement • Measures of Academic Progress • QualityCore • PSAT • AP exam
Nontest measures	• Attendance • Graduation • Course completion	• Course pass rate • Dropout
Student learning objectives	• Student learning objectives	• Student growth objectives
Evaluation and performance	• School performance • School evaluation • Teacher performance	• Teacher evaluation • Principal evaluation

Source: Authors.

Table A2. Queries by alternative growth measure type

Category	Alternative assessment	Nontest measure	Student learning objective
Student growth	✔	✔	
Alternative assessments	✔		
Nontest measures		✔	
Student learning objectives			✔
Evaluation and performance		✔	✔

Source: Authors.

performance of cohorts or grade levels were excluded. Growth models include simple gains analyses, residual gains, multivariate linear regressions, and quantile regressions, among others.

Search results

Table A3 counts the total, relevant, and useful citations identified in each query of the database searches (EBSCO Education Research Complete, EconLit, and Google Scholar). Of the 307 studies identified in the database search, only 7 contained relevant information (2.3 percent). Google Scholar does not have the same search options as the library databases—for example, it does not allow word proximity searches or Boolean queries with more than 256 characters. As a result, the Google Scholar queries resulted in large numbers of citations, many of them irrelevant. The searches identified 592 citations for alternative assessments and 108 for SLOs. Of the 73 studies mentioning use of alternative measures in growth models, 36 contained relevant information (5.1 percent of all studies identified).[3]

The results of the library database searches combined with a search of websites of districts, states, research organizations, and foundations known to be using alternative growth models (based on sites identified in the literature and through professional contacts)

Table A3. Database search results

Query	Number of citations	Number mentioning use of alternative measures in growth models	Number relevant for review	Percent relevant for review
Library database search				
Alternative assessments	23	8	2	8.7
Nontest measures	21	5	2	9.5
Student learning objectives	263	16	3	1.1
Google Scholar search				
Alternative assessments	592	55	25	4.2
Student learning objectives	108	18	11	10.2

Source: Authors.

identified 91 citations that provided useful information on the implementation or the statistical properties of alternative student growth models (table A4).

More than two-thirds of the 91 citations identified were descriptive materials published by districts, states, research organizations, foundations, test vendors, value-added vendors, or periodicals. These citations were identified largely through websites of districts or states, websites of organizations such as the Teacher Incentive Fund Community and the Value-Added Research Center, and Google searches. The descriptive materials provided general information on how alternative measures are incorporated into student growth models and how they are used for teacher evaluation, school improvement, compensation, and other purposes. Some materials also included information on how alternative assessments are created or selected, benefits and drawbacks of various approaches, and reactions from teachers on the use of such measures. These descriptive materials did not include information on the statistical properties of the measures.

Of the 91 citations identified, 30 provided information on the statistical properties of alternative student growth models or SLOs (see table A4), most of them including some measures of reliability or correlation with other measures. Of these 30 citations, 8 were published in peer-reviewed journals, and 22 were published by research organizations, by foundations, or in working papers.

Table A4. Summary of citations identified

Type of citation	Number of citations identified
Descriptive materials on implementation	
Published by districts or states	30
Published by other entities (research organizations, foundations, test vendors, value-added vendors, periodicals)	31
Research on statistical properties of measures	
Published in peer-reviewed publications	8
Published by research organizations, by foundations, or in working papers	22

Source: Authors.

One reason for the scant evidence on alternative measures in student growth models is that the measures have yet to be widely implemented. Table A5 counts the locations identified as piloting or using alternative measures in student growth models or SLOs. Less than half the locations (42 percent) have been using such measures for five years or more (since 2008/09 or earlier). Fourteen locations are in their first three years of implementation. Appendixes B and C present more details on the locations using these types of measures and the research that has been conducted in some of these locations.

Table A5. Summary of findings on implementation in districts and states

Measure or objective	Number of locations piloting or using measures	Number of locations using measures for five or more years (2008/09 or earlier)
Alternative measures in student growth models		
State-mandated secondary-level end-of-course assessments	7	4
Local assessments	5	2
Alternative standardized assessments	19	10
Attendance	3	2
Other nontest outcomes	1	0
Total[a]	23	12
Student learning objectives		
Based on state standardized assessments	9	0
Based on alternative assessments	12	3
Total	13	3

a. The categories within alternative measures and student learning objectives are not mutually exclusive. Because some locations use more than one type of measure, the totals are less than the sum of locations using each type of measure.

Source: Authors.

Appendix B. Results of the literature search for alternative student outcomes in statistical growth models

The studies contributing to the findings on the use of alternative assessments in student growth models are summarized in table B1. These are the key studies that included the most information on statistical properties of measures and were the most relevant for educators and policymakers. All these studies met the following criteria:

- They use statistical models of student growth that rely on longitudinal data on individual students, thereby controlling for students' prior achievement.
- With one exception, they estimate value-added at the teacher level.

Table B1 also indicates whether each study passed through a formal or informal peer-review process (to the extent that could be determined). Formal peer reviews by academic journals are distinguished from other kinds, including reviews by government agencies, internal organizations, dissertation committees, and other researchers.

Table B2 summarizes the research on the statistical properties of specific alternative student outcome measures. This table may be of special interest to readers who want further information on specific assessments or to determine which assessments have more evidence and which have less.

Table B3 lists the locations that use alternative growth models in school or teacher evaluations, as identified in the studies. This table may be of special interest to readers interested in learning more about implementation across the country or who want to contact districts or states to learn more about specific implementation issues.

Table B4 summarizes key information on all studies and documents identified in the literature search that relate to alternative growth models. It includes the studies excluded from the analysis and descriptive documentation on the use of these measures in districts and states.

Table B1. Key studies of alternative measures in growth models

Study	Location	Grades and subjects	Alternative growth measure	Correlation with other performance measures	Reliability and other statistical properties
Biancarosa, G., Bryk, A., & Dexter, E. R. (2010)	17 schools in 8 Eastern states	K–2, reading	Teacher-level value-added estimated with Dynamic Indicators of Basic Early Literacy Skills and Terra Nova (scaled together)	—	*Reliability* Correlates teacher value-added across four years: 0.38–0.71.
Bill & Melinda Gates Foundation (2010)	Charlotte-Mecklenburg, NC; Dallas, TX; Denver, CO; Hillsborough County, FL; New York City	4–8, math and reading	Teacher-level value-added estimated with Balanced Assessment in Mathematics and Stanford 9 Open-Ended Reading Assessment	Correlates value-added estimated with alternative assessments and value-added estimated with state assessments with different group of students. Estimates true correlation by calculating persistent variance in each measure: • M: 0.54 • R: 0.37 (all districts); 0.59 (without New York City, which switched state assessments in 2009/10) Correlates value-added estimated with alternative assessments and student survey measures from different group of students: • M: 0.11 (sum of all measures); 0.15 (control and challenge measures) • R: 0.06 (sum of all measures); 0.10 (control and challenge measures)	*Reliability* Correlates value-added estimated within teacher across different classes of students in same year (one year of assessment data): • M: 0.23 • R: 0.35 *Distribution of teacher effects* • M: SD = 0.26; interdecile range = 0.63 • R: SD = 0.34; interdecile range = 0.79
Bill & Melinda Gates Foundation (2012)[a]	Charlotte-Mecklenburg, NC; Dallas, TX; Denver, CO; Hillsborough County, FL; New York City; Memphis, TN	4–8, math and reading	Teacher-level value-added estimated with Balanced Assessment in Mathematics and Stanford 9 Open-Ended Reading Assessment	Correlates value-added estimated with alternative assessments and value-added estimated with state assessments: • M: 0.45 • R: 0.46 Calculates difference in alternative assessment value-added between teachers ranked in top and bottom quartiles on various measures: • Teacher observation instruments: 0.05–0.11 (M); 0.10–0.16 (R) • Student survey: 0.06 (M); 0.05 (R) • State assessment value-added model: 0.11 (M); 0.09 (R) • Combination of measures: 0.08–0.13 (M); 0.10–0.15 (R)	*Distribution of teacher effects* Interquartile range (in months of learning): • M: 4.5 months • R: 4.8 months

(continued)

Table B1. Key studies of alternative measures in growth models *(continued)*

Study	Location	Grades and subjects	Alternative growth measure	Correlation with other performance measures	Reliability and other statistical properties
Corcoran, S. P., Jennings, J. L., & Beveridge, A. A. (2011)[a]	Houston, TX	4–5, math and reading	Teacher-level value-added estimated with Stanford Achievement Test	Correlates value-added estimated with Stanford and value-added with state assessment with same students: • M: 0.59 • R: 0.50 Estimates the proportion of value-added that persists from one year to the next: • M, Stanford: 0.61 • M, state: 0.40 • R, Stanford: 0.61 • R, state: 0.42	*Distribution of teacher effects* • M: SD of stable teacher effects = 0.22; SD of teacher-by-year effects = 0.25 • R: SD of stable teacher effects = 0.17; SD of teacher-by-year effects = 0.20
Gray, J. J. (2010)	Large Midwestern district	7–8, English, math, and communication	Teacher-level value-added estimated with Measures of Academic Progress	Predicts value-added estimated with alternative assessment using principal rankings of teachers. Principal rankings are a statistically significant predictor of math value-added. They do not significantly predict English value-added.	—
Harris, D. N., & Sass, T. R. (2012)[a]	Midsize Florida district	2–10, math and reading	Teacher-level value-added estimated with Stanford Achievement Test	Correlates value-added estimated with Stanford and average state assessment scale score gain: • M: 0.19–0.27 • R: –0.40 to 0.13 Correlates value-added estimated with Stanford and principal ratings: • Ratings of teacher characteristics: 0.19–0.34 (M); 0.20–0.45 (R) • Ratings of teacher's ability to raise test scores: 0.27 (M); 0.14 (R) Percentage of teachers ranked in bottom 30 percent on Stanford value-added and principal ratings: 65 percent (M); 54 percent (R) Percentage of teachers ranked in top 30 percent on Stanford value-added and principal ratings: 16 percent (M); 21 percent (R)	—

(continued)

Table B1. Key studies of alternative measures in growth models *(continued)*

Study	Location	Grades and subjects	Alternative growth measure	Correlation with other performance measures	Reliability and other statistical properties
Jackson, C. K. (2012)	North Carolina	9, algebra and English	Teacher-level value-added estimated with end-of-course exams	—	*Reliability* Covariance between years: • M: 0.13 (without school track effects); 0.01 (with school track effects) • R: 0.01 (without and with school track effects) *Distribution of teacher effects* Without school track effects: • M: SD = 0.23 (0.12 attributed to true teacher quality) • R: SD = 0.16 (0.06 attributed to true teacher quality) With school track effects: • M: SD = 0.14 (0.08 attributed to true teacher quality) • R: SD = 0.08 (0.04 attributed to true teacher quality)
Johnson, M., Lipscomb, S., Gill, B., Booker, K., & Bruch, J. (2012)[a]	Pittsburgh, PA	1–12 (high school only for core course pass rate)	Teacher-level value-added estimated with CBAs; school-level value-added estimated with attendance and core course pass rate	—	*Reliability* Percentage of teachers/schools distinguishable from average: Teacher effects (average across grades): • M, CBA: 38 percent • R, CBA: 19 percent • Science, CBA: 46 percent • Social studies, CBA: 47 percent School effects: • Attendance, grades 1–3: 5 percent • Attendance, grades 4–8: 33 percent • Attendance, grades 9–12: 58 percent • Core course pass rate: 75 percent *Distribution of teacher and school effects* Difference between 90th percentile teacher and mean (*z*-score units): Teacher effects (average across grades): • M, CBA: 0.52 • R, CBA: 0.26 • Science, CBA: 0.48 • Social studies, CBA: 0.47 School effects: • Attendance, grades 1–3: 0.05 • Attendance, grades 4–8: 0.10 • Attendance, grades 9–12: 0.21 • Core course pass rate: 0.12

(continued)

Table B1. Key studies of alternative measures in growth models *(continued)*

Study	Location	Grades and subjects	Alternative growth measure	Correlation with other performance measures	Reliability and other statistical properties
Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013)[a]	Charlotte-Mecklenburg, NC; Dallas, TX; Denver, CO; Hillsborough County, FL; New York City; Memphis, TN	4–8, math and reading; 9, math, English language arts, and biology	Teacher-level value-added estimated with Balanced Assessment in Mathematics, Stanford 9 Open-Ended Reading Assessment, and ACT QualityCore	Grades 4–8: A teacher predicted to raise student achievement on state assessment by 1 SD is predicted to raise student achievement on alternative assessments by 0.66 SD (math and reading analyzed together). Grade 9: A teacher predicted to raise student achievement on state assessment by 1 SD is predicted to raise student achievement on alternative assessments by: • M: 0.58 • R: 1.05 • Biology: 0.96 • All three subjects: 0.63–0.83	—
Lipscomb, S., Gill, B., Booker, K., & Johnson, M. (2010)[a]	Pittsburgh, PA	6–12 (all measures); 4–5 (attendance)	Locally developed CBAs, SRI, PSAT (teacher and school); student attendance and credit accumulation (school only)	Correlates value-added estimated with alternative measures with value-added estimated with state assessment in same subject: Teacher effects: • R, CBA: 0.31 • R, SRI: 0.25 • Science, CBA: 0.05–0.24 School effects: • M, CBA: 0.55 • R, CBA: 0.71 • R, SRI: 0.42 • Science, CBA: –0.06 to 0.30 • History, CBA: 0.35 (compared to PSSA reading) • Attendance: –0.36 to –0.06 (range across all PSSA subjects) • Credits earned: 0.17–0.32 (range across all PSSA subjects)	*Reliability* Correlates value-added within school across years: • M, CBA, grade 9: 0.43 • R, CBA, grade 9: 0.86 • Science, CBA, grade 9: 0.71 • Attendance, grades 9 and 10: 0.13–0.57 • Credits earned, grades 9 and 10: 0.72–0.87 Correlates value-added within teacher across years: • M, CBA, grade 9: 0.14 • M, PSAT, grade 11: 0.05 • R, CBA, grade 9: –0.24 • R, PSAT, grade 11: 0.37 • Writing, PSAT, grade 11: 0.73 • Science, CBA, grade 9: 0.77 • Civics, CBA, grade 9: 0.65

(continued)

Table B1. Key studies of alternative measures in growth models *(continued)*

Study	Location	Grades and subjects	Alternative growth measure	Correlation with other performance measures	Reliability and other statistical properties
McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009)[b]	Five districts in Florida: Dade, Duval, Hillsborough, Orange, and Palm Beach	3–8, math	Teacher-level value-added estimated with Stanford Achievement Test	—	*Reliability* Correlates value-added within teacher across years (varies across county and specification): • Elementary: 0.16–0.46 • Middle: 0.28–0.61 Year-to-year quintile rankings: • Elementary: 32–39 percent of those in top quintile in first year were also in top quintile in second year; 24–29 percent of those in top quintile in first year were in bottom two quintiles in second year • Middle: 28–38 percent in top quintile in both years; 22–28 percent of those in top quintile in first year were in bottom two quintiles in second year
Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013)[a]	Charlotte-Mecklenburg, NC; Dallas, TX; Denver, CO; Hillsborough County, FL; New York City; Memphis, TN	4–8, math and reading	Teacher-level value-added estimated with Balanced Assessment in Mathematics and Stanford 9 Open-Ended Reading Assessment	Correlates the stable components in value-added estimated with alternative assessments and other measures: • State assessment value-added model: 0.39–0.43 (M); 0.41–0.54 (R) • Teacher observation instrument: 0.10–0.42 (M); 0.30–0.35 (R) • Student survey: 0.19–0.32 (M); 0.16–0.33 (R) Correlates value-added estimated with alternative assessments with composite measure that includes state assessment value-added, observations, and student surveys: • M: 0.31 (elementary); 0.45 (middle) • R: 0.38 (elementary); 0.35 (middle)	*Reliability* Estimates reliability based on section-to-section variability and variability from aggregating measures across students: • M: 0.33 (elementary); 0.69 (middle) • R: 0.29 (elementary); 0.80 (middle)

(continued)

Table B1. Key studies of alternative measures in growth models *(continued)*

Study	Location	Grades and subjects	Alternative growth measure	Correlation with other performance measures	Reliability and other statistical properties
Papay, J. P. (2011)[b]	Large Northeastern district	3–5, reading	Teacher-level value-added estimated with Stanford Achievement Test and SRI	Correlates value-added estimated with alternative assessments and state assessment with same students: • Stanford and state: 0.16 • SRI and state: 0.44 • Stanford and SRI: 0.27 53.1 percent of teachers are ranked in top quartile on both the SRI and state assessments; 24.8 percent of those ranked in top quartile on state assessment are ranked in bottom two quintiles on SRI.	*Distribution of teacher effects* • Stanford: SD = 0.05 • SRI: SD = 0.21
Sass, T. R. (2008)	San Diego, CA; four districts in Florida: Duval, Hillsborough, Orange, and Palm Beach	Elementary and middle school (Florida); high school (San Diego)	Teacher-level value-added estimated with Stanford Achievement Test	Correlation of quintile ranking on Stanford and state assessment (Hillsborough only) = 0.48. 43 percent of teachers in top quintile on one assessment ranked in top quintile on the other; 13 percent of teachers ranked in top quintile on Stanford ranked in bottom two quintiles on state assessment.	*Reliability* Correlates value-added estimates within teacher across years (varies across county and year): • Elementary: 0.08–0.36 • Middle: 0.18–0.38 Year-to-year quintile rankings: 22–31 percent in top quintile in both years; 21–31 percent of those in top quintile in first year in bottom two quintiles in second year.

— is not available; CBA is curriculum-based assessment; M is math assessment; PSAT is Preliminary Scholastic Aptitude Test; PSSA is Pennsylvania System of School Assessment; R is reading assessment; SD is standard deviation; SRI is Scholastic Reading Inventory.

a. Study was peer reviewed by experts other than the authors but not subjected to formal academic journal peer review. Includes reviews by government agencies, internal organizations, dissertation committees, and other researchers.

b. Study underwent formal external peer review for journal publication.

Source: Authors' analysis based on search criteria described in appendix A.

Table B2. Research on alternative student outcomes in growth models

Measure	Growth model reliability data available?[a]	Growth model correlation with other measures available?	Other growth model statistics available?	Locations studied
National standardized tests				
ACT QualityCore exams		✔	✔	Charlotte-Mecklenburg, NC Dallas, TX Denver, CO Hillsborough County, FL Memphis, TN New York City 23 schools in Midwest
Balanced Assessment in Mathematics	✔	✔		Charlotte-Mecklenburg, NC Dallas, TX Denver, CO Hillsborough County, FL Memphis, TN New York City
DIBELS and Terra Nova (scaled together)	✔		✔	17 schools in 8 Eastern states
Measures of Academic Progress		✔	✔	Unknown large district
PSAT	✔	✔	✔	Pittsburgh Public Schools
Stanford Achievement Test	✔	✔	✔	Charlotte-Mecklenburg, NC Dallas, TX Dade County, Duval County, Hillsborough County, Orange County, and Palm Beach County, FL Denver, CO Houston, TX Memphis, TN New York City San Diego, CA Unknown large districts
Scholastic Reading Inventory		✔		Unknown large urban district
State or local assessments				
End-of-course exams in 17 states and the District of Columbia (included in EPIC charter school study)			✔	California Colorado District of Columbia Florida Georgia Hawaii Illinois Indiana Louisiana Massachusetts Michigan Minnesota Missouri New Mexico New York Ohio Pennsylvania Texas

(continued)

Table B2. Research on alternative student outcomes in growth models *(continued)*

Measure	Growth model reliability data available?[a]	Growth model correlation with other measures available?	Other growth model statistics available?	Locations studied
Hillsborough composite measure[b]			✔	Hillsborough County, FL
Pittsburgh course-based assessments		✔	✔	Pittsburgh Public Schools
North Carolina end-of-course exams	✔			North Carolina
Tennessee end-of-course exams			✔	Memphis Public Schools
Nontest measures				
Attendance	✔	✔	✔	Pittsburgh Public Schools
Credits earned	✔	✔	✔	Pittsburgh Public Schools
Core courses passed			✔	Pittsburgh Public Schools
Holding power			✔	Pittsburgh Public Schools

DIBELS is Dynamic Indicators of Basic Early Literacy Skills; PSAT is Preliminary Scholastic Aptitude Test.

a. Reliability is defined as stability of value-added estimates over time.

b. Based on Stanford Achievement Test, Advanced Placement/International Baccalaureate exams, and local end-of-course assessments.

Source: Authors' analysis based on search criteria described in appendix A.

Table B3. Implementation of alternative student outcomes in growth models

Location	State-mandated secondary level end-of-course assessments	Local assessments	Alternative standardized assessments	Nontest measure	Teacher- or school-level measure?	First year implemented
States						
North Carolina	✔				Both	2005/06
Ohio			✔		Teacher	2011/12
Tennessee	✔		✔		Teacher	1992/93
Districts/selected schools						
Atlanta Public Schools, GA	✔				Both	2010/11
Charlotte-Mecklenburg Schools, NC[a]	✔	✔	✔		Teacher	2009/10
Chicago Public Schools, IL			✔		School	1996/97
Dallas Independent School District, TX		✔	✔	✔	Both	1995/96
Denver Public Schools, CO			✔		Teacher	2009/10
EPIC charter schools in 20 states	✔				Both	2006/07
Hillsborough County Public Schools, FL		✔	✔		Teacher	2009/10
Houston, TX			✔		Teacher	2006/07
Little Rock School District, AR			✔		Teacher	2004/05
Memphis City Schools, TN	✔		✔		School	2006/07
Meridian School District, ID			✔		Teacher	1999/2000
Milwaukee Public Schools, WI		✔	✔	✔	School	2005/06
New York City Public Schools, NY			✔		Teacher	2009/10
Pittsburgh Public Schools, PA		✔	✔	✔	Both	2010/11
Racine Unified School District, WI			✔		Teacher	2005/06
SIATech Charter Network			✔		School	Unknown
SOAR districts in Ohio[b]			✔		Teacher	2008/09
TIF charter schools in Detroit, MI[c]			✔		Both	2011/12
Tulsa Public Schools, OK	✔				Both	2009/10
Winston-Salem/Forsyth County, NC[d]			✔		Both	2011/12

SOAR is Schools' Online Achievement Reports; TIF is Teacher Incentive Fund.

a. Nearly 3,000 teacher volunteers in six districts—Charlotte-Mecklenburg, Dallas, Denver, Hillsborough County, Memphis, and New York City—participated in the Bill & Melinda Gates Foundation's Measures of Effective Teaching project in 2009/10 and 2010/11.

b. In Ohio's SOAR districts 133 high schools are participating in a Batelle for Kids high school value-added initiative.

c. The Michigan Association of Public School Academies was awarded a five-year TIF3 grant from the U.S. Department of Education in 2010. The grant funds the implementation of teacher effectiveness measures and performance-based compensation in 20 charter schools in Detroit.

d. The Iowa Test of Basic Skills is used to measure student growth in 16 schools in Winston-Salem/Forsyth County that are participating in the TIF3 grant from the U.S. Department of Education.

Source: Authors' analysis based on search criteria described in appendix A.

Table B4. All relevant studies identified on alternative measures in growth models

Study	Location	Grades and subjects	Alternative measure used in growth model	Implementation stage of growth measures	Includes evidence on statistical properties?	Notes
Battelle for Kids. (n.d.)	Ohio	9–12, nine core subject areas	ACT QualityCore end-of-course exams	Implemented in 36 schools in 2008; expanded to 44 schools by 2011	No	Frequently asked questions and implementation information on Ohio High School Value-added Initiative, which is designed to identify best practices and foster professional development
Biancarosa, G., Bryk, A., & Dexter, E. R. (2010)[a]	17 schools in 8 Eastern states	K–2	DIBELS and Terra Nova scaled together using Rasch modeling	Used to evaluate impact of school reform model	Reliability and other statistical properties	Estimates teacher and school value-added during multiyear implementation of a schoolwide reform model; correlates teacher and school value-added estimates over time
Bill & Melinda Gates Foundation (2010)	Charlotte-Mecklenburg, NC; Dallas, TX; Denver, CO; Hillsborough County, FL; New York City	4–8, math and reading	Balanced Assessment in Mathematics; Stanford 9 Open-Ended Reading Assessment	Implementation began in 2009/10 as part of MET	Correlation with other measures and other statistical properties	Correlates value-added estimated with alternative assessments with value-added estimated with state assessment
Bill & Melinda Gates Foundation (2012)[b]	Charlotte-Mecklenburg, NC; Dallas, TX; Denver, CO; Hillsborough County, FL; New York City; Memphis, TN	4–8, math and reading	Balanced Assessment in Mathematics; Stanford 9 Open-Ended Reading Assessment	Implementation began in 2009/10 as part of MET	Correlation with other measures and other statistical properties	Correlates value-added estimated with alternative assessments with value-added estimated with state test
Bill & Melinda Gates Foundation (2013)[b]	Charlotte-Mecklenburg, NC; Dallas, TX; Denver, CO; Hillsborough County, FL; New York City; Memphis, TN	4–8, math and reading	Balanced Assessment in Mathematics; Stanford 9 Open-Ended Reading Assessment	Implementation began in 2009/10 as part of MET	Correlation with other measures	Correlates value-added estimated with alternative assessments with composite measure based on state assessment value-added, teacher observations, and student surveys
Burnett, A., Cushing, E., & Bivona, L. (2012)[a]	—	—	End-of-course exams	—	No	Describes strengths and weaknesses of various alternative growth measures, including end-of-course exams and student learning objectives

(continued)

Table B4. All relevant studies identified on alternative measures in growth models *(continued)*

Study	Location	Grades and subjects	Alternative measure used in growth model	Implementation stage of growth measures	Includes evidence on statistical properties?	Notes
Cantrell, S. M. (2012)[a]	Charlotte-Mecklenburg, NC; Dallas, TX; Denver, CO; Hillsborough County, FL; New York City; Memphis, TN	4–8, math and reading	Balanced Assessment in Mathematics; Stanford 9 Open-Ended Reading Assessment	Implementation began in 2009/10 as part of MET	Other statistical properties	Describes predictive power of alternative assessments and state assessments in terms of future student achievement
Clark, L. (2002)[a]	Meridian, ID	3–12	Measures of Academic Progress	Implemented to track student learning in 1999/2000	No	Describes teacher and administrator attitudes toward use of Measures of Academic Progress
Community Training and Assistance Center (2013)[b]	Charlotte-Mecklenburg, NC	9–12	North Carolina end-of-course exams	Implementation began in 2009/10 as part of TIF3 grant	No	Describes implementation of value-added model based on state end-of-course exams in high school
Corcoran, S. P., Jennings, J. L., & Beveridge, A. A. (2011)[b]	Houston, TX	4–5, math and reading	Stanford Achievement Test	Used only for research	Correlation with other measures and other statistical properties	Correlates value-added estimated with Stanford Achievement Test with value-added estimated with state test; estimates proportion of value-added that persists from one year to the next on each assessment
Curtis, R. (2012a)	Hillsborough County, FL	1–12, all subjects	End-of-course exams, Stanford Achievement Test, Advanced Placement/ International Baccalaureate exams, and other assessments	Implementation began in 2010/11; teachers received first rating in fall 2012	Other statistical properties	Describes implementation of new teacher evaluation system and presents distribution of value-added scores
Curtis, R. (2012b)	Charlotte-Mecklenburg, NC	Unknown	End-of-course exams	Implementation began in 2010/11 and assessment development is ongoing	No	Describes implementation of new end-of-course exams to include more teachers in value-added models
Dawson, L., Mallory, K., & Johnson, K. (2011)[a]	SIATech charter schools	9–12	Renaissance Learning Reading and Mathematics assessments	Implemented to track school-level growth; not used for teacher evaluation	No	Describes use of alternative assessments to compare growth across schools in network of dropout recovery charter schools

(continued)

Table B4. All relevant studies identified on alternative measures in growth models *(continued)*

Study	Location	Grades and subjects	Alternative measure used in growth model	Implementation stage of growth measures	Includes evidence on statistical properties?	Notes
Goe, L., & Holdheide, L. (2011)[a]	—	—	—	—	No	General implementation guidance on evaluating teachers in nontested positions, with examples from Hillsborough County, Austin, and Delaware
Gray, J. J. (2010)[b]	Large Midwestern district	7–8, math, English, and communication	Measures of Academic Progress	Used only for research	Correlation with other measures and other statistical properties	Correlates value-added with average gain scores and principal ratings
Harris, D. N., & Sass, T. R. (2012)[b]	Midsize Florida district	2–10, math and reading	Stanford Achievement Test (FCAT-NRT)	Used only for research	Correlation with other measures	Correlates value-added with principal ratings
Harris, D. N., & Sass, T. R. (2010)	Midsize Florida district	2–10, math and reading	Stanford Achievement Test (FCAT-NRT)	Used only for research	Correlation with other measures	Correlates teacher value-added with principal ratings
Heck, R. H. (2009)[a]	Large Western district	3–5, math and reading	Stanford Achievement Test	Used only for research	Other statistical properties	Estimates effect of prior and current teacher value-added and school-level value-added on achievement
Jackson, C. K. (2012)	North Carolina	9, algebra and English	North Carolina end-of-course exams	Used only for research	Reliability	Estimates predictive power of high school teacher value-added considering selection into tracks and unobserved track-level treatments
Johnson, M., Lipscomb, S., Gill, B., Booker, K., & Bruch, J. (2012)[b]	Pittsburgh, PA	1–12	Locally developed curriculum-based assessments (teacher and school); PSAT, student attendance, core course pass rate, holding power (school only)	Implemented as part of teacher evaluation and compensation system in 2009/10; adjusted assessments and models in 2010/11	Correlation with other measures and other statistical properties	Correlates value-added composites (which use state tests and alternative measures) with Pennsylvania Value-Added Assessment System, which uses state tests and a different value-added model; reports distribution of value-added based on each measure by grade and subject
Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013)[b]	Charlotte-Mecklenburg, NC; Dallas, TX; Denver, CO; Hillsborough County, FL; New York City; Memphis, TN	4–8, math and reading	Balanced Assessment in Mathematics, Stanford 9 Open-Ended Reading Assessment, and ACT QualityCore	Implementation began in 2009/10 as part of MET	Correlation with other measures	Predicts value-added estimated with alternative assessments based on value-added with state assessments

(continued)

Table B4. All relevant studies identified on alternative measures in growth models *(continued)*

Study	Location	Grades and subjects	Alternative measure used in growth model	Implementation stage of growth measures	Includes evidence on statistical properties?	Notes
Keller, B. (2006)	Houston, TX	Unknown	Stanford Achievement Test	Implemented as part of compensation system in 2006/07	No	Describes implementation of teacher compensation system
Koedel, C., & Betts, J. (2010)[a]	San Diego, CA	4, math	Stanford Achievement Test	Used only for research	Other statistical properties	Estimates teacher-level value-added with simulated ceiling effects
Lipscomb, S., Gill, B., Booker, K., & Johnson, M. (2010)[b]	Pittsburgh, PA	1–12	Locally developed curriculum-based assessments, Scholastic Reading Inventory, PSAT (teacher and school); student attendance and credit accumulation (school only)	Implemented as part of teacher and school evaluation system in 2009/10	Reliability, correlation with other measures, and other statistical properties	Correlates value-added estimated with alternative assessments with nontest outcomes estimated with state assessment, reports year-to-year reliability for subset of measures, reports distribution of effects for all measures
Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007)[a]	Large district	6–8, math	Stanford Achievement Test	Used only for research	Other statistical properties	Correlates teacher value-added across subscores of assessment and model specifications
Lombardi, K. A. (2011)[b]	Urban district in Kansas	3–5	Measures of Academic Progress	Used in district to track student progress and informally evaluate teachers	No	Describes use of Measures of Academic Progress for informal teacher evaluation and includes teacher and principal attitudes about assessment
Marietta, G. (n.d.)	Hillsborough County, FL	K–12, all subjects	End-of-course exams	Implemented as part of compensation system in 2006/07; coverage expanded in subsequent years	No	Describes district's experience creating local assessments for all grades and subjects to be used in a value-added model
McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009)[a]	Five districts in Florida: Dade, Duval, Hillsborough, Orange, and Palm Beach	3–8, math	Stanford Achievement Test (FCAT-NRT)	Used only for research	Reliability	Describes year-to-year correlation of value-added for alternative assessment and state assessment

(continued)

Table B4. All relevant studies identified on alternative measures in growth models *(continued)*

Study	Location	Grades and subjects	Alternative measure used in growth model	Implementation stage of growth measures	Includes evidence on statistical properties?	Notes
Meyer, R. H., Carl, B., & Cheng, H. E. (2010)	Milwaukee, WI	9	District grade 9 quarterly benchmark exams, attendance	Implemented as pilot in 2009/10	No	Describes pilot of using grade 9 exams and attendance in a value-added model
Michigan Association of Public School Academies (2012)	Detroit, MI	2–12	Scantron Performance Series, ACT Series	Implemented as part of compensation system in 2011/12 in charter schools participating in TIF3 grant	No	Implementation handbook describes use of alternative assessments in compensation system
Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013)[b]	Charlotte-Mecklenburg, NC; Dallas, TX; Denver, CO; Hillsborough County, FL; New York City; Memphis, TN	4–8, math and reading	Balanced Assessment in Mathematics; Stanford 9 Open-Ended Reading Assessment	Implementation began in 2009/10 as part of MET	Correlation with other measures and reliability	Correlates value-added estimated with alternative assessments with value-added estimated with state assessment, teacher observations, student surveys, and composite measure
NCTechNews (2012)	North Carolina	K–12	—	Implemented as part of teacher evaluation system in 2012/13	No	Press release on use of SAS Education Value-Added Assessment System K–12 in North Carolina
Ohio Department of Education (2012a,b)	Ohio	K–12	Commercially available assessments and locally developed assessments (approved by state)	Implemented as part of teacher evaluation system in 2012/13	No	Describes state teacher evaluation system, which includes value-added measures based on alternative assessments
Papay, J. P. (2011)[a]	Large Northeastern district	3–5, reading	Stanford Achievement Test, Scholastic Reading Inventory	Used only for research	Correlation with other measures and other statistical properties	Correlates value-added estimated with alternative assessments with value-added estimated with state assessment; examines different testing time periods
Potamites, L., Booker, K., Chaplin, D., & Isenberg, E. (2009)[b]	EPIC charter school consortium (145 charter schools in 17 states and the District of Columbia)	9–12, core subjects	End-of-year assessments (varies by state)	Implemented as part of school evaluation system in 2006/07; part of teacher evaluation system in 2007/08	Other statistical properties	Presents results of value-added analysis conducted for EPIC charter school consortium evaluation system

(continued)

Table B4. All relevant studies identified on alternative measures in growth models *(continued)*

Study	Location	Grades and subjects	Alternative measure used in growth model	Implementation stage of growth measures	Includes evidence on statistical properties?	Notes
Potamites, L., Chaplin, D., Isenberg, E., & Booker, K. (2009)[b]	Memphis, TN	9–12, algebra, English, and biology	Tennessee end-of-course exams	Implemented as part of evaluation system in 2008/09	Other statistical properties	Presents results of value-added analysis conducted for Memphis evaluation system
Prince, C. D., Shuermann, P. J., Guthrie, J. W., Witham, P. J., Milanowski, A. T., & Thorn, C. A. (2009)	—	Nontested grades and subjects	End-of-course exams, DIBELS, Measures of Academic Progress, ACT, SAT, and other assessments	—	No	Provides general guidance on evaluating teachers in nontested positions with examples and lessons from states
Sass, T. R. (2008)	San Diego, CA; four districts in Florida: Duval, Hillsborough, Orange, and Palm Beach	Elementary and middle (Florida); high school (San Diego)	Stanford Achievement Test	Used only for research	Reliability and correlation with other measures	Summarizes data similar to that presented in McCaffrey et al. (2009) and Koedel and Betts (2007)
Sass, T. R. (2011)	Florida	4–10, math and reading	Stanford Achievement Test (FCAT-NRT)	Used only for research	Correlation with other measures and other statistical properties	Includes only new teachers; compares teachers from different certification paths
SCORE (2012)	Tennessee	9–12	End-of-course exams and ACT Series	Used as part of teacher evaluation system starting in 2010/11	No	Describes state teacher evaluation system, which includes high school assessments
Steele, J. L., Hamilton, L. S., & Stecher, B. M. (2010)[b]	—	—	End-of-course exams	—	No	Describes examples and benefits and drawbacks of various types of assessments used in growth models
Tulsa Public Schools (2011)	Tulsa, OK	9–12	End-of-instruction exams	Used for instructional purposes starting in 2009/10, will be part of evaluation system in 2013/14	No	Press release on 2010/11 value-added reports

(continued)

Table B4. All relevant studies identified on alternative measures in growth models *(continued)*

Study	Location	Grades and subjects	Alternative measure used in growth model	Implementation stage of growth measures	Includes evidence on statistical properties?	Notes
Webster, W. J., & Mendro, R. L. (1997)	Dallas, TX	K–12	Iowa Test of Basic Skills, attendance	Implemented for school evaluation starting in 1992/93 and for teacher evaluation in 1995/96	No	Describes use of alternative assessments and attendance in school and teacher evaluation systems, including benefits and drawbacks
Winters, M., Greene, J. P., Ritter, G., & Marsh, R. (2008)[a]	Little Rock, AR	K–5, math, reading, and language	Iowa Test of Basic Skills	Implemented as part of compensation pilot in five schools in district over three years	No	Describes implementation in three schools that implemented compensation pilot for the first time in the third year of the pilot

— is not available; DIBELS is Dynamic Indicators of Basic Early Literacy Skills; FCAT-NRT is Florida Comprehensive Assessment Test Norm Referenced Test; MET is Bill & Melinda Gates Foundation's Measures of Effective Teaching project; SAT is Scholastic Aptitude Test; TIF is Teacher Incentive Fund.

a. Study underwent formal external peer review process for journal publication.

b. Study was peer reviewed by experts other than the authors but not subjected to formal academic journal peer review. Includes reviews by government agencies, internal organizations, dissertation committees, and other researchers.

Source: Authors' analysis based on search criteria described in appendix A.

Appendix C. Results of the literature search for student learning objectives

The studies contributing to the findings on student learning objectives (SLOs) are summarized in tables C1 and C2. Table C1 includes all SLO studies with data on reliability, validity, or the percentage of teachers meeting SLOs. Only seven studies included such information. Although evidence is limited on the statistical properties of SLOs, this table provides readers with the locations for which this information is available. Table C2 summarizes the findings from the key studies that include implementation lessons based on data collected from teachers or districts. These key studies exclude reports that did not document the systematic collection of empirical data. This table may be of special interest to readers interested in educators' experiences implementing SLOs.

Table C3 includes information on where SLOs are being used for teacher evaluation and indicates which type of assessments are used in each location. Because SLOs are implemented and used for evaluations in various ways, the table also includes a brief description of the unique features of the SLOs in each location. This table may be of special interest to readers seeking to learn more about where and how SLOs have been implemented.

Table C4 summarizes the data and methods used in each key implementation study in table C2. It includes data collection methods, sample sizes, and response rates for each study. This information may be of special interest to readers seeking to learn more about how the implementation data were collected and the extent to which the data reflect the experiences of participating educators.

Table C5 includes all the studies and documents identified in the literature search that describe SLO statistical properties or implementation. It includes documents that were excluded from table C2 because they were not based on data systematically collected from teachers or districts. It also includes material published by districts and states with guidance on how to create SLOs and incorporate them in evaluation systems. These studies may be of interest to readers seeking to learn more about how SLOs have been implemented across the country.

Table C1. Statistical properties of student learning objectives

Study	Location	Correlation with other performance measures	Reliability and other statistical properties
Community Training and Assistance Center (2004)	Denver, CO	Reports mean normal curve equivalent on state assessment and Iowa Test of Basic Skills for teachers meeting zero, one, and two student learning objectives (SLOs): • Elementary: teachers meeting two SLOs generally have higher mean test scores than teachers meeting one or zero • Middle and high school: less definitive pattern of association between meeting SLOs and mean test scores	89–93 percent of teachers met at least one SLO each year of the four-year pilot of Denver's ProComp professional development and compensation system. Classroom teachers who participated in the pilot for longer periods met SLOs at higher rates: • One year of participation: 89 percent met at least one SLO • Two years: 93 percent • Three years: 94 percent • Four years: 98 percent Rigor of SLOs improved over time: 0 percent of classroom teacher SLOs were rated at highest level of rigor in first year; 21 percent rated at highest level in fourth year
Community Training and Assistance Center (2013)	Charlotte-Mecklenburg, NC	Analyzes relationship between SLO attainment and student achievement on state tests: • Math: positive significant association in years 1 and 2 • Reading: some positive significant association in all years (elementary only in year 2, grade 6 only in year 3) Teachers who receive bonus based on value-added are more likely to have high-quality SLOs than teachers who do not receive a value-added model bonus (statistically significant in year 2)	Percentage of SLOs met: • 2008/09: 61 percent • 2009/10: 81 percent • 2010/11: 55 percent Percentage of SLOs met in 2010/11 by teacher's years of experience in SLO initiative: • One year: 49 percent • Two years: 51 percent • Three years: 64 percent
Goldhaber, D., & Walch, J. (2011)	Denver, CO	Examines percentage of teachers receiving ProComp SLO awards by quintile of value-added. Finds slightly higher percentages of awards for teachers in top two value-added model quintiles than in bottom two quintiles: • Math: SLO awards to 44 percent of teachers in top two quintiles against 37 percent in bottom two quintiles • Reading: SLO awards to 42 percent of teachers in top two quintiles against 37 percent in bottom two quintiles Percentage of teachers who earned incentive based on SLOs that are above/below mean value-added based on state test: • Math: 30 percent (above); 31 percent (below) • Reading: 41 percent (above); 34 percent (below) Logistic regression finds small but statistically significant relationship between earning an SLO award and value-added	Percentage of ProComp teachers earning an incentive tied to meeting SLOs: • 2006/07: 77 percent • 2007/08: 82 percent • 2008/09: 85 percent • 2009/10: 84 percent

(continued)

Table C1. Statistical properties of student learning objectives *(continued)*

Study	Location	Correlation with other performance measures	Reliability and other statistical properties
Proctor, D., Walters, B., Reichardt, R., Goldhaber, D., & Walch, J. (2011)	Denver, CO	Correlates percentage of SLOs met at school level with school-level growth indicator: r = .096 (p = .29)	Percentage of participating ProComp teachers earning an incentive tied to meeting SLOs[a]: · 2006/07: 71 percent · 2007/08: 76 percent · 2008/09: 80 percent · 2009/10: 80 percent
Schmitt, L., & Ibanez, N. (n.d.)	Austin, TX	Teachers that met at least one SLO generally have higher net growth on the Texas state test than teachers that met no SLOs (overall and for novice teachers)	—
Tennessee Department of Education (2012)	Tennessee	—	65 percent of teachers received the highest score possible on teacher-selected portion of the evaluation system
Terry, B. D. (2008)	Austin, TX	—	Percentage of teachers meeting SLOs in first pilot year: · Meeting at least one SLO: 83 percent · Meeting both SLOs: 64 percent

— is not available.

a. Goldhaber and Walch (2011) and Proctor, Walters, Reichardt, Goldhaber, and Walch (2011) report slightly different percentages of Denver teachers earning SLO incentives in each year. They both use administrative data from the district, but it is possible that their samples vary. For example, Goldhaber and Walch (2011) cite issues with linking data across data sources due to masked teacher IDs. This may have resulted in a different analytic sample than that used by Proctor et al. (2011).

Source: Authors' analysis based on search criteria described in appendix A.

Table C2. Implementation findings for student learning objectives

Study	Location	Implementation lessons and teacher attitudes	Use in evaluation or compensation
Austin Independent School District (2012)	Austin, TX	• Staff attitudes were more positive when the principal was supportive of the program. • Elementary teachers participating in the student learning objective (SLO) process were more likely than comparison teachers to discuss professional development needs and goals, discuss assessment data for individual students, set SLOs for groups of students, and group students based on learning needs. There were no significant differences between participants and comparisons at the middle or high school levels. • Some teachers were frustrated that student mobility, dropouts, and attendance made it difficult to meet SLOs.	SLOs are part of REACH, the district's strategic compensation program. Teachers earn $1,000 for each SLO met ($1,500 at high-need campuses).
Burns, S. F., Gardner, C. D., & Meeuwsen, J. (2009)	Austin, TX	• 67 percent of teachers agree that SLOs are a positive change. • More than 80 percent viewed SLOs as a highly or moderately important component of REACH. • 66 percent disagree that SLOs are a good measure of effective teaching. • 61 percent disagree that REACH distinguishes effective and ineffective teachers. • 67 percent report that REACH is fair to teachers. • 31 percent of those who did not meet SLOs report that REACH is fair.	SLOs are part of REACH, the district's strategic compensation program. Teachers earn $1,000 for each SLO met ($1,500 at high-need campuses).
Community Training and Assistance Center (2004)	Denver, CO	• Teachers in pilot reported having better access to student data and that they use the data more effectively. • Most teachers do not attribute core classroom instructional changes to participation in pilot. • Most teachers reported that cooperation among teachers improved or stayed the same since pilot began. • Teachers found the SLO process complex even though they had already been doing other forms of objective setting in prior years. • Teachers learned how to use student achievement data and set reasonable goals. • Greater instructional support and feedback are needed.	SLOs were tied to compensation in pilot. In the pilot's first year teachers earned $500 per SLO met; in the second year they earned $750 per SLO met.
Community Training and Assistance Center (2013)	Charlotte-Mecklenburg, NC	• Teachers reported that they used SLOs to improve student learning and that the SLO process made them more focused and knowledgeable about students' strengths and needs. • Teachers valued data analysis, planning, and instructional elements of SLOs. • Improving timeliness and availability of data improves quality of SLOs. • Issues with software used to document SLOs was a distraction to participants.	SLOs are part of the district's Teacher Incentive Fund performance-based compensation program. Teachers earn $1,400 per SLO met ($1,000 per SLO in year 3).

(continued)

Table C2. Implementation findings for student learning objectives *(continued)*

Study	Location	Implementation lessons and teacher attitudes	Use in evaluation or compensation
Proctor, D., Walters, B., Reichardt, R., Goldhaber, D., & Walch, J. (2011)	Denver, CO	• More than 60 percent of teachers and principals reported that SLOs improve instructional practices. • About 50 percent of teachers report that SLOs impact professional growth. • Close to 60 percent of teachers and principals believe that SLOs will increase student achievement. • 34 percent of principals report that SLOs make administrative work less difficult. • 28 percent of principals report that SLOs make administrative work more difficult. • Perceptions of SLO implementation and experience range from neutral to slightly positive. • Some teachers believe that SLOs were not implemented consistently across schools.	SLOs are a component of the district's compensation system. Teachers that meet both SLOs receive a 1 percent base-building incentive. Teachers that meet one SLO receive a 1 percent non-base-building incentive.
Reform Support Network (2012)	—	• Creating a theory of action is essential for defining the purpose and intended outcomes of SLOs. • Providing training on how to help teachers create assessments and SLOs is important. • Training should be accompanied by tools, including rubrics, examples, and timelines. • Training can be provided in creative ways, such as through webinars, embedded in other types of professional development, and with school-based SLO facilitators. • Quality assurance can be achieved with automated data systems and audits.	Based on examples from multiple locations.
Schmitt, L., & Ibanez, N. (n.d.)	Austin, TX	• Some teachers reported that they spend a lot of time on SLOs, but the SLOs are not always related to instructional improvement. • SLO training tends to be focused more on mechanics of the process than on instruction.	SLOs are a component of the district's REACH compensation system.
Tennessee Department of Education (2012)	Tennessee	• Assessment choices were too often based on teacher and principal beliefs about which assessment will provide the highest score. • Many teachers did not see the benefits of the system, and SLOs are not viewed as drivers of effective teaching. • Teachers viewed the SLO component as one of the least effective components of the evaluation system. This is because measures were selected inconsistently and the teachers often did not receive data back until the following year. • Teachers reported more intentional use of student data, more schoolwide collaboration, and new kinds of conversations around instruction and outcomes.	SLOs are a component of the state's teacher evaluation system. They account for 15 percent of the evaluation rating.

(continued)

Table C2. Implementation findings for student learning objectives *(continued)*

Study	Location	Implementation lessons and teacher attitudes	Use in evaluation or compensation
TNTP (2012)	Indiana	• Teachers felt that the SLO process was time-consuming, particularly creating and updating assessments. • Obtaining prior-year student data to serve as a baseline was challenging. • Majority of teachers believe that SLOs should accompany other measures of student learning in evaluation system. • Identifying or creating assessments should be done prior to or at the beginning of the school year. • Technology solutions are needed for storing student learning data.	SLOs are a component of RISE, the state's evaluation system. SLOs account for 10–20 percent of evaluation, depending on availability of value-added data for teachers.

Source: Authors' analysis based on search criteria described in appendix A. For details on the data and methods used in each key implementation study, see table C4.

Table C3. Implementation of student learning objectives

Location	Outcome measures incorporated in student learning objectives		Unique features of student learning objectives
	State accountability assessments?	Alternative assessments?	
State			
Delaware	✔	✔	—
Georgia	✔	✔	State centralized system, with each district leading the development of 2–3 student learning objectives (SLOs) and assessments. SLOs are created for courses, not for individual teachers. The state approves all SLOs.
Indiana	✔	✔	Districts have the option of adopting state SLO system. In the first year of implementation teachers set two SLOs based on one class (preferably classes without state assessment data available). In the future teachers will set SLOs for all classes.
New York	✔	✔	State provides list of approved assessments and guidance on which assessments should be used by different categories of teachers.
Ohio	✔	✔	State provides guidance on hierarchy of assessments and how to create tiered targets. Teachers or teams of teachers set SLOs, and SLO evaluators approve them.
Rhode Island	✔	✔	State provides guidance on how to select assessments and incorporate them into SLOs. General education and special education teachers work together to align SLOs.
Tennessee	✔	✔	Teachers select goal at beginning of year; goal could be based on state or commercially available assessment, graduation rate, promotion rate, or completion of advanced coursework.
District			
Denver Public Schools, CO		✔	Teachers design two SLOs each year based on commercially available assessments, district assessments, or teacher-developed assessments. SLOs are approved by principals and submitted to an online forum.
District of Columbia Public Schools		✔	District provides guidance on suggested assessments and performance levels for each grade and subject. SLOs can be based on multiple assessments.
Austin Independent School District, TX		✔	Teachers design two SLOs each year: one for whole class, and one targeted at student subgroups. SLOs are designed by teachers, approved by principals, and evaluated for rigor by district team.
Jeffco Public Schools, CO	✔		SLOs are based on state exams until alternative assessments are identified.
Charlotte-Mecklenburg Schools, NC		✔	The district piloted SLOs as part of a Teacher Incentive Fund grant, but decided to build them into standard teaching practice rather than incorporating them into a districtwide compensation system.
New Haven Public Schools, CT	✔	✔	—

— is not available.

Source: Authors' analysis based on search criteria described in appendix A.

Table C4. Data and methods in key student learning objective implementation studies

Study	Source of information	Data collection method	Sample size	Response rates
Austin Independent School District (2012)	Teachers, assistant principals, other school staff	Focus groups (REACH schools) Teacher surveys (REACH and comparison schools)	Focus groups: 240 Teacher survey: —	—
Burns, S. F., Gardner, C. D., & Meeuwsen, J. (2009)	Teachers, principals	Surveys (teachers) Interviews	Survey: 449 Teacher interviews: 34 Principal interviews: 10	Survey: 71 percent Teacher interviews: 49 percent Principal interviews: —
Community Training and Assistance Center (2004)	Teachers, principals, parents in pilot and control schools; board members, association leaders, district administrators, external community members	Surveys (teachers, principals, parents) Interviews Focus groups Observations	Pilot school educator survey: average 359 a year (1,436 over four years) Control school educator survey: average 284 a year (851 over three years) Parent survey: average 194 a year (583 over three years) Educator interviews: 370 Other interviews: 245	Pilot educator survey: 53–83 percent Control educator survey: 33–39 percent Parent survey: 4–10 percent Interviews: —
Community Training and Assistance Center (2013)	Teachers, principals, parents, other stakeholders	Focus groups Interviews Teacher and principal web surveys Parent phone surveys	Focus groups and interviews: 934 (209 teachers) Teacher and principal survey: 23,707 Parent survey: 2,026	Teacher and principal survey: 50.6 percent overall (94–96 percent for teachers)
Proctor, D., Walters, B., Reichardt, R., Goldhaber, D., & Walch, J. (2011)	Teachers, administrators, teacher trainees, district staff, other stakeholders	Surveys (teachers, administrators, and teacher trainees) Interviews Focus groups	Teacher survey: 2,985 Administrator survey: 169 Teacher trainee survey: 350 Teacher interviews: 250 Administrator interviews: 36 District staff interviews: 17 Stakeholder interviews: 13	Teacher survey: 61 percent Administrator survey: 72 percent Teacher trainee survey: 20 percent Interviews: —
Reform Support Network (2012)	State and district officials	Interviews	—	—
Schmitt, L., & Ibanez, N. (n.d.)	Teachers	Surveys	—	—
Tennessee Department of Education (2012)	Teachers, administrators, district evaluators, other stakeholders	Surveys Meetings Emails Interviews	Teacher survey: ~16,000 Administrator survey: ~1,000 Meetings with educators: 7,500 Interviews with district evaluators and school administrators: 42	—
TNTP (2012)	Teachers, evaluators, central office staff	Surveys (teachers and evaluators) Interviews Focus groups	—	—

— is not available.

Source: Authors' analysis based on search criteria described in appendix A.

Table C5. All relevant studies identified on student learning objectives

Study	Location	Notes about source
Austin Independent School District (2012)	Austin, TX	Presents results of educator focus groups and surveys on student learning objectives (SLOs) and other aspects of the REACH strategic compensation system
Bagshaw, T., & Holdheide, L. (2010)	na	Describes implementation lessons, with a focus on evaluating special education teachers
Brodsky, A., DeCesare, D., & Kramer-Wine, J. (2010)[a]	Austin, TX	Summarizes several teacher compensation systems, including Austin's REACH program in its first year of implementation
Burnett, A., Cushing, E., & Bivona, L. (2012)[b]	na	Describes strengths and weaknesses of various types of alternative growth measures, including end-of-course exams and SLOs
Burns, S. F., Gardner, C. D., & Meeuwsen, J. (2009)[a]	Austin, TX	Focuses on pilot phase of implementation, describes teacher attitudes based on surveys, interviews, and document reviews
Casson, C., & Good, B. (2012)	Denver, CO	Describes challenges with implementing SLOs and potential solutions
Center for Educator Compensation Reform (n.d.)	Charlotte-Mecklenburg, NC	Describes implementation of SLOs as part of Teacher Incentive Fund compensation system
Community Training and Assistance Center (2004)[b]	Denver, CO	Evaluates SLO implementation during four-year pilot phase; includes distribution of SLOs met, analysis of rigor of SLOs, correlations with student achievement, and teacher attitudes and experiences
Community Training and Assistance Center (2013)[b]	Charlotte-Mecklenburg, NC	Evaluates SLO implementation during a three-year Teacher Incentive Fund grant period; includes distribution of SLOs met, analysis of rigor of SLOs, analysis of relationships between SLO attainment/quality and student achievement, teacher attitudes, and experiences
Curtis, R. (2012b)	Charlotte-Mecklenburg, NC	Describes implementation of SLOs as part of a new teacher evaluation and compensation system
District of Columbia Public Schools (2011a)	Washington, DC	Gives district guidance on teacher evaluation system, which incorporates SLOs for nontested teachers
District of Columbia Public Schools (2011b)	Washington, DC	Gives district guidance for developing SLOs and using them to measure teacher performance
EducationCounsel (2012, February)	na	Describes benefits and challenges of implementing SLOs in several locations
Fulbeck, E. S., & Farley, A. N. (2012, November)	Denver, CO	Examines relationship between teacher attitudes and behaviors; attitude data are from teacher survey
Georgia Department of Education (2011, 2012)	Georgia	Gives state guidance for developing SLOs and using them to measure teacher performance
Goe, L., & Holdheide, L. (2011)[a]	na	Gives general implementation guidance on evaluating teachers in nontested positions, with examples from Hillsborough County, Austin, and Delaware
Goldhaber, D., & Walch, J. (2011)[a]	Denver, CO	Describes implementation of strategic compensation system and distribution of teachers receiving incentives based on SLOs
Indiana Department of Education (n.d.)	Indiana	Gives state guidance for developing SLOs and using them to measure teacher performance
Lachlan-Haché, L. (2012)	na	Presents lessons from examples in Austin, Denver, Rhode Island, Indiana, and Ohio
Lachlan-Haché, L., Cushing, E., & Bivona, L. (2012)[b]	na	Describes SLO challenges and solutions from examples in Austin, Rhode Island, Georgia, Ohio, Indiana, and New York
Miller, A. (2012)	Indiana	Describes incorporation of SLOs into teacher evaluation system
New Haven Public Schools (2010)	New Haven, CT	Gives district guidance for selecting measures for SLOs
New York State Education Department (2012a,b)	New York	Gives state guidance for developing SLOs and using them to measure teacher performance
Ohio Department of Education (2012a)	Ohio	Gives state guidance for developing SLOs and using them to measure teacher performance

(continued)

Table C5. All relevant studies identified on student learning objectives *(continued)*

Study	Location	Notes about source
Ohio Department of Education (2012b)	Ohio	Gives state guidance on teacher evaluation system, which incorporates SLOs
Proctor, D., Walters, B., Reichardt, R., Goldhaber, D., & Walch, J. (2011)	Denver, CO	Reports distribution of types of SLOs, correlation with schoolwide growth measure, rates of meeting SLOs, and teacher and principal attitudes
Race to the Top Technical Assistance Network (n.d.)	na	Gives general guidance about benefits and challenges of implementing SLOs, using examples from several locations
Reform Support Network (2012)	na	Describes implementation lessons from five districts/states
Reform Support Network (n.d.)	na	Describes aspects of high-quality SLOs and general SLO benefits, challenges, and solutions
Sawchuk, S. (2011)	na	Gives general guidance on evaluating teachers in nontested positions, with examples from Charlotte-Mecklenburg, Austin, and Denver
Schmitt, L., & Ibanez, N. (n.d.)	Austin, TX	Studies relationship between meeting SLOs and net growth on state assessment; also includes some implementation information on SLO training and teacher attitudes
Steele, J. L., Hamilton, L. S., & Stecher, B. (2010)[b]	na	Describes implementation of SLOs in Denver and Washington, DC
Tennessee Department of Education (2012)	Tennessee	Describes implementation of first year of new evaluation system, which includes SLOs for teachers in nontested positions; includes description of SLO types, rates of meeting SLOs, and implementation challenges
Terry, B. D. (2008)	Austin, TX	Describes strategic compensation system, which incorporates SLOs and results from pilot in nine schools in 2007/08
TNTP (2011)	Washington, DC	Describes implementation of SLOs in Washington, DC
TNTP (2012)	Indiana	Describes implementation lessons and teacher attitudes from first year of implementation
White, S. (2012)	Georgia	Describes phased implementation of SLOs

na is not applicable.

a. Study underwent formal external peer review process for journal publication.

b. Study was peer reviewed by experts other than the authors but not subjected to formal academic journal peer review. Includes reviews by government agencies, internal organizations, dissertation committees, and other researchers.

Source: Authors' analysis based on search criteria described in appendix A.

Notes

1. There is debate about the extent to which standard value-added models produce valid measures of teacher effectiveness (see, for example, Rothstein, 2010; Goldhaber & Chaplin, 2012), but available evidence suggests that any bias in standard value-added models is likely to be small (Kane, McCaffrey, Miller, & Staiger, 2013; Kane & Staiger, 2008).
2. For example, the following query was used for nontest measures: (("value-added" OR "value added" OR "student growth" OR "residual gain") N10 ("attendance" OR "graduation" OR "course completion" OR "course pass rate" OR "dropout")) AND ("school performance" OR "school evaluation" OR "teacher performance" OR "teacher evaluation" OR "principal evaluation"). The Boolean term "N10" indicates that one of the terms in the first string must be found within ten words of a term in the second string.
3. A Google Scholar search was not conducted for nontest measures because the query could not be narrowed sufficiently.

References

Austin Independent School District. (2012). *AISD REACH Program Update, 2010–2011: Participant feedback.* (Department of Research and Evaluation Report 10.86 RB). Austin: Author. Retrieved December 13, 2012, from http://www.austinisd.org/sites/default/files/dre-reports/rb/10.86RB_AISD_Reach_Participant_Feedback_2010–2011_0.pdf

Bagshaw, T., & Holdheide, L. (2010). *Special education teacher evaluation: Issues and answers.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved December 13, 2012, from http://www.ode.state.or.us/initiatives/elearning/nasdse/jan12presentationslides.pdf

Battelle for Kids. (n.d.). *Frequently asked questions, Ohio value-added high schools.* Retrieved October 23, 2012, from http://portal.battelleforkids.org/Ohio/SOAR/High_School_Initiative/FAQs.html?sflang=en

Biancarosa, G., Bryk, A., & Dexter, E. R. (2010). Assessing the value-added effects of literacy collaborative professional development on student learning. *The Elementary School Journal, 111*(1), 7–34.

Bill & Melinda Gates Foundation. (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching project.* (MET Project Research Paper.) Seattle: Author. Retrieved October 23, 2012, from http://www.gatesfoundation.org/college-ready-education/Documents/preliminary-findings-research-paper.pdf

Bill & Melinda Gates Foundation. (2012). *Gathering feedback for teaching.* (MET Project Research Paper). Seattle: Author. Retrieved October 23, 2012, from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Bill & Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching.* (MET Project Policy and Practitioner Brief). Seattle: Author. Retrieved January 9, 2013, from http://metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf

Brodsky, A., DeCesare, D., & Kramer-Wine, J. (2010). Design and implementation considerations for alternative teacher compensation programs. *Theory Into Practice, 49*(3), 213–222.

Burnett, A., Cushing, E., & Bivona, L. (2012). *Use of multiple measures for performance-based compensation.* Washington, DC: Center for Educator Compensation Reform. Retrieved December 14, 2012, from http://0-cecr.ed.gov.opac.acc.msmc.edu/pdfs/CECR_MultipleMeasures.pdf

Burns, S. F., Gardner C. D., & Meeuwsen, J. (2009). *An evaluation of teacher and principal experiences during the pilot phase of AISD REACH: A strategic compensation initiative.* Nashville, TN: Peabody College at Vanderbilt University. Retrieved December 13, 2012, from http://discoverarchive.vanderbilt.edu/jspui/bitstream/1803/3379/1/BurnsGardnerMeeuwsen%20Capstone%20May%202009.pdf

Cantrell, S. M. (2012). The measures of effective teaching project: An experiment to build evidence and trust. *Education Finance and Policy, 7*(2), 203–218.

Casson, C., & Good, B. (2012). *Student growth (learning) objectives in Denver: Past experiences and future.* Denver: Denver Public Schools. Retrieved December 13, 2012, from https://www.tifcommunity.org/sites/default/files/Session%205B-%20Casson%20&%20Good-%20Student%20Growth%20Objectives%20in%20Denver.pdf

Center for Educator Compensation Reform. (n.d.). *Community Training and Assistance Center and the Charlotte-Mecklenburg Schools Leadership for Educator's Advanced Performance.* Washington, DC: Author. Retrieved December 14, 2012, from http://www.cecr.ed.gov/initiatives/profiles/pdfs/CommunityTrainingandAssistanceCenter.pdf

Clark. L. (2002). The hard drive to student growth. *School Administrator, 4*(59), 24–27.

Community Training and Assistance Center. (2004). *Catalyst for change: Pay for performance in Denver: Final report.* Boston: Author. Retrieved May 4, 2012, from http://www.ctacusa.com/PDFs/Rpt-CatalystChangeFull-2004.pdf

Community Training and Assistance Center. (2013). *It's more than the money.* Boston: Author. Retrieved February 27, 2013, from http://www.ctacusa.com/PDFs/MoreThanMoney-report.pdf

Corcoran, S. P., Jennings, J. L., & Beveridge, A. A. (2011). *Teacher effectiveness and high- and low-stakes tests.* (Working paper). New York: New York University. Retrieved October 23, 2012, from https://files.nyu.edu/sc129/public/papers/corcoran_jennings_beveridge_2011_wkg_teacher_effects.pdf

Curtis, R. (2012a). *Building it together: The design and implementation of Hillsborough County Public Schools' teacher evaluation system.* Washington, DC: The Aspen Institute. Retrieved December 17, 2012, from http://www.aspendrl.org/portal/browse/DocumentDetail?documentId=1068&download

Curtis, R. (2012b). *Putting the pieces in place: Charlotte-Mecklenburg Public Schools' teacher evaluation system.* Washington, DC: The Aspen Institute. Retrieved December 17, 2012, from http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/ed-CharlotteREP4.pdf

Dawson, L., Mallory, K., & Johnson, K. (2011). A focus on individual student growth. *Leadership, 40*(3), 22–26.

District of Columbia Public Schools. (2011a). *IMPACT: The DCPS effectiveness assessment system for school-based personnel, 2011–2012.* Washington, DC: Author. Retrieved October 23, 2012, from http://www.dc.gov/DCPS/Files/downloads/TEACHING%20&%20LEARNING/IMPACT/IMPACT%20Guidebooks%202010–2011/Impact%202011%20Group%202-Aug11.pdf

District of Columbia Public Schools. (2011b). *Teacher-assessed student achievement data (TAS) guidance*. Washington, DC: Author. Retrieved October 23, 2012, from http://www.isbe.net/peac/pdf/DCPS_SLO_guidance_022412.pdf

EducationCounsel. (2012, February). *Student learning objectives: An emerging framework*. Presented to the Illinois Performance Evaluation Advisory Council, Springfield, IL. Retrieved October 23, 2012, from http://www.isbe.state.il.us/peac/pdf/slo_emerg_frmwk_pres_02–12.pdf

Fulbeck, E. S., & Farley, A. N. (2012, November). *Denver ProComp: Teachers' attitudes and behaviors*. Presented at the annual conference of the Association for Public Policy & Analysis, Baltimore, MD. Retrieved December 13, 2012, from http://appam.confex.com/appam/2012/webprogram/Paper3443.html

Georgia Department of Education. (2011). *Teacher keys evaluation system handbook: Pilot, January–May 2012*. Atlanta: Author. Retrieved October 23, 2012, from http://legisweb.state.wy.us/InterimCommittee/2012/TKESHandbook.pdf

Georgia Department of Education. (2012). *Student learning objectives operations manual*. Atlanta: Author. Retrieved October 23, 2012, from http://www.doe.k12.ga.us/School-Improvement/Teacher-and-Leader-Effectiveness/Documents/SLO%20Manual.pdf

Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D., & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added*. Washington, DC: Brookings Institution.

Goe, L., & Holdheide, L. (2011). *Measuring teachers' contributions to student learning: Growth for nontested grades and subjects*. (Research & Policy Brief). Washington, DC: National Comprehensive Center for Teacher Quality.

Goldhaber, D., & Chaplin, D. (2012). *Assessing the Rothstein test: Does it really show teacher value-added models are biased?* (Working paper). Washington, DC: Mathematica Policy Research. Retrieved December 17, 2012, from http://mathematica-mpr.com/publications/pdfs/education/rothstein_wp.pdf

Goldhaber, D., & Hansen, M. (2010). *Is it just a bad class? Assessing the stability of measured teacher performance*. (CEDR Working Paper 2010-3). Seattle: Center for Education Data and Research.

Goldhaber, D., & Walch, J. (2011). Strategic pay reform: A student outcomes-based evaluation of Denver's ProComp teacher pay initiative. *Economics of Education Review, 31*(6), 1067–1083.

Gray, J. J. (2010). Are principals good at identifying effective teachers? A comparison of teachers' principal ratings and residual gains on standardized tests. (Doctoral dissertation, Kansas University, 2010). Retrieved December 13, 2012, from http://kuscholarworks.ku.edu/dspace/bitstream/1808/6765/1/Gray_ku_0099D_10785_DATA_1.pdf

Harris, D. N., & Sass, T. R. (2012). *Skills, productivity and the evaluation of teacher performance*. Retrieved December 13, 2012, from http://www2.gsu.edu/~tsass/IES%20Harris%20Sass%20Principal%20Eval%2056%20clean%20-%20AEJ%20Format.pdf

Harris, D. N., & Sass, T. R. (2010). *What makes for a good teacher and who can tell?* Revised. Retrieved December 14, 2012, from http://myweb.fsu.edu/tsass/Papers/IES%20Harris%20Sass%20Principal%20Eval%2034.pdf

Heck, R. H. (2009). Teacher effectiveness and student achievement: Investigating a multilevel cross-classified model. *Journal of Educational Administration, 47*(2), 227–249.

Indiana Department of Education. (n.d.). *RISE Evaluation and Development System: Student learning objectives handbook*. Retrieved October 23, 2012, from http://www.riseindiana.org/sites/default/files/files/Student%20Learning/Student%20Learning%20Objectives%20Handbook%201%200%20FINAL.pdf

Jackson, C. K. (2012). *Teacher quality at the high-school level: The importance of accounting for tracks*. (Working Paper No. 17722). Cambridge, MA: National Bureau of Economic Research. Retrieved December 13, 2012, from http://www.nber.org/papers/w17722

Johnson, M., Lipscomb, S., Gill, B., Booker, K., & Bruch, J. (2012). *Value-added models for the Pittsburgh Public Schools*. (Report to Pittsburgh Public Schools). Cambridge, MA: Mathematica Policy Research.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers?* (MET Project Research Paper). Seattle: Bill & Melinda Gates Foundation. Retrieved January 9, 2013, from http://metproject.org/downloads/MET_Validating_Using_Random_Assignment_Research_Paper.pdf

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation*. (Working Paper No. 14607). Cambridge, MA: National Bureau of Economic Research. Retrieved December 17, 2012, from http://www.nber.org/papers/w14607

Keller, B. (2006). Test-tied bonuses to take effect in Houston. *Education Week, 25*(19), 5–12.

Koedel, C., & Betts, J. (2010). Value-added to what? How a ceiling in the testing instrument influences value-added estimation. *Education Finance and Policy, 5*(1), 54–81.

Lachlan-Haché, L. (2012). *SLO implementation considerations: Communication, training, and assessment literacy*. Santa Monica, CA: Teacher Incentive Fund. Retrieved December 13, 2012, from https://www.tifcommunity.org/sites/default/files/Session%205A-%20Lachlan%20Hache%20SLO%20Implementation%20Considerations.pdf

Lachlan-Haché, L., Cushing, E., & Bivona, L. (2012). *Student learning objectives: Benefits, challenges, and solutions*. Washington, DC: American Institutes for Research. Retrieved December 13, 2012, from http://educatortalent.airprojects.org/inc/docs/SLOs_Benefits_Challenges_Solutions.pdf

Lipscomb, S., Gill, B., Booker, K., & Johnson, M. (2010). *Estimating teacher and school effectiveness in Pittsburgh: Value-added modeling and results, second edition.* (Report to Pittsburgh Public Schools). Cambridge, MA: Mathematica Policy Research.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47–67.

Lombardi, K. A. (2011). *Testing the limits: The purposes and effects of additional, external elementary mathematics assessment.* Lawrence, KS: Kansas University. Retrieved December 14, 2012, from http://kuscholarworks.ku.edu/dspace/bitstream/1808/8149/1/Lombardi_ku_0099D_11347_DATA_1.pdf

Marietta, G. (n.d.). *Multiple measures of teacher effectiveness in Hillsborough County Public Schools: Implementing value-added measures*. Seattle: Bill & Melinda Gates Foundation. Retrieved October 23, 2012, from http://www.fadss.org/_docs/_content/eet/caseStudyHCPS/ImplementingValue-addedMeasuresHCPS.pdf

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572–606.

Meyer, R. H., Carl, B., & Cheng, H. E. (2010). *Accountability and performance in secondary education in Milwaukee Public Schools*. (The Senior Urban Education Research Fellowship Series, vol. II). Washington, DC: The Council of the Great City Schools. Retrieved December 14, 2012, from http://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/85/Milwaukee_FellowReport2010.pdf

Michigan Association of Public School Academies. (2012). *TEAMS (Teacher Excellence and Academic Milestones for Students) playbook*. Lansing, MI: Michigan Association of Public School Academies. Retrieved October 23, 2012, from http://www.teams.charterschooldissemination.org/images/stories/TEAMS%20Playbook/TEAMS_Playbook_5–7-12.pdf

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A composite estimator of effective teaching.* (MET Project Research Paper). Seattle: Bill & Melinda Gates Foundation. Retrieved January 9, 2013, from http://www.metproject.org/downloads/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf

Miller, A. (2012). *RISE student learning objectives*. Indianapolis: Teacher Incentive Fund, Indiana Department of Education. Retrieved December 14, 2012, from https://www.tifcommunity.org/index.php?q=content/session-7-combining-measures

NCTechNews. (2012). *North Carolina to use SAS EVAAS for K-12 in teacher, principal evaluations*. June 29. Retrieved October 30, 2012, from http://nctechnews.com/2012/06/29/saas/north-carolina-to-use-sas%c2%ae-evaas%c2%ae-for-k-12-in-teacher-principal-evaluations/7471/

New Haven Public Schools. (2010). *NHPS teacher evaluation and development: Student learning goals*. Retrieved October 30, 2012, from http://www.nhps.net/sites/default/files/8__Student_Learning_Growth_-_Goal_Setting_Introduction_100825.pdf

New York State Education Department. (2012a). *Guidance on the New York State District-wide growth goal-setting process: Student learning objectives*. Albany: Author. Retrieved October 23, 2012, from http://engageny.org/sites/default/files/resource/attachments/slo-guidance.pdf

New York State Education Department. (2012b). *New York State district-wide growth goal setting process: Student learning objectives. Road map for districts*. Retrieved October 23, 2012, from http://engageny.org/sites/default/files/resource/attachments/slo-roadmap.pdf

Ohio Department of Education. (2012a). *A guide to using student learning objectives as a locally-determined measure of student growth*. Retrieved October 23, 2012, from http://www.ode.state.oh.us/GD/DocumentManagement/DocumentDownload.aspx?DocumentID=134104

Ohio Department of Education. (2012b). *Ohio teacher evaluation system model*. Retrieved October 30, 2012, from http://www.ode.state.oh.us/GD/DocumentManagement/DocumentDownload.aspx?DocumentID=128955

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163–193.

Potamites, L., Booker, K., Chaplin, D., & Isenberg, E. (2009). *Measuring school and teacher effectiveness in the EPIC Charter School Consortium—year 2: Final report*. Washington, DC: Mathematica Policy Research.

Potamites, L., Chaplin, D., Isenberg, E., & Booker, K. (2009). *Measuring school effectiveness in Memphis—year 2: Final report*. Washington, DC: Mathematica Policy Research.

Prince, C. D., Shuermann, P. J., Guthrie, J. W., Witham, P. J., Milanowski, A. T., & Thorn, C. A. (2009). *The other 69 percent: Fairly rewarding the performance of teachers of nontested subjects and grades*. Washington, DC: Center for Educator Compensation Reform. Retrieved December 14, 2012, from http://www.cecr.ed.gov/guides/other69Percent.pdf

Proctor, D., Walters, B., Reichardt, R., Goldhaber, D., & Walch, J. (2011). *Making a difference in education reform: ProComp external evaluation report 2006–2010*. Prepared for the Denver Public Schools. Denver: The Evaluation Center, University of Colorado.

Race to the Top Technical Assistance Network. (n.d.). *Measuring student growth for teachers in nontested grades and subjects: A primer*. Retrieved December 13, 2012, from http://www.swcompcenter.org/educator_effectiveness2/NTS__PRIMER_FINAL.pdf

Reform Support Network. (2012). *Lessons learned around developing and implementing student learning objectives*. Prepared for the Maryland Department of Education.

Retrieved December 13, 2012, from http://www.marylandeducators.org/uploadedFiles/MSEA_Content/Your_Profession/School_Quality/Student%20Learning%20Objectives.pdf

Reform Support Network. (n.d.). *Targeting growth: Using student learning objectives as a measure of educator effectiveness*. Retrieved December 13, 2012, from http://ok.gov/sde/sites/ok.gov.sde/files/TLE-NonTestedGradesSub-SLOTargeting.pdf

Rothstein, J. (2010). Teacher quality in education production: Tracking, decay, and student achievement. *Quarterly Journal of Economics, 125*(1), 175–214.

Sass, T. R. (2008). *The stability of value-added measures of teacher quality and implications for teacher compensation policy*. (National Center for Analysis of Longitudinal Data in Education Research Brief 4). Washington, DC: The Urban Institute. Retrieved December 13, 2012, from http://www.urban.org/UploadedPDF/1001266_stabilityofvalue.pdf

Sass, T. R. (2011). *Certification requirements and teacher quality: A comparison of alternative routes to teaching*. Atlanta: Georgia State University. Retrieved December 13, 2012, from http://www2.gsu.edu/~tsass/Alternative%20Certification%20and%20Teacher%20Quality%2011.pdf

Sawchuk, S. (2011). Wanted: Ways to measure most teachers. *Education Week, 30*(19), 1–15.

Schmitt, L., & Ibanez, N. (n.d.). *AISD REACH program update: 2009–2010 Texas Assessment of Knowledge and Skills (TAKS) results and student learning objectives (SLOs)*. (Department of Program Evaluation Publication 09.83 RB). Austin: Austin Independent School District.

Schochet, P., & Chiang, H. S. (2010). *Error rates in measuring teacher and school performance based on student test score gains*. (NCEE 2010–4004). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

SCORE. (2012). *TVAAS: An introduction to value-added in Tennessee*. (Taking Note: Examining Key Education Reform Ideas in Tennessee, June 2012). Nashville, TN: Author. Retrieved December 13, 2012, from http://www.tnscore.org/wp-content/uploads/2010/12/SCORE_TVAAS4.pdf

Steele, J. L., Hamilton, L. S., & Stecher, B. M. (2010). *Incorporating student performance measures into teacher evaluation systems*. Santa Monica, CA: RAND Corporation. Retrieved December 14, 2012, from http://www.rand.org/content/dam/rand/pubs/technical_reports/2010/RAND_TR917.pdf

Sturman, M. C., Cheramie, R. A., & and Cashen, L. H. (2005). The impact of job complexity and performance measurement on the temporal consistency, stability, and test-retest reliability of employee job performance ratings. *Journal of Applied Psychology, 90*(2), 269–283.

Tennessee Department of Education. (2012). *Teacher evaluation in Tennessee: A report on year 1 implementation*. Nashville, TN: Author. Retrieved December 14, 2012, from http://www.tn.gov/education/doc/yr_1_tchr_eval_rpt.pdf

Terry, B. D. (2008). *Paying for results: Examining incentive pay in Texas schools*. Austin: Texas Public Policy Foundation. Retrieved December 13, 2012, from http://www.broadeducation.org/asset/1128-paying%20for%20results.pdf

TNTP. (2011). *DCPS-TNTP IMPACT conference: Focus on measuring student achievement*. Retrieved October 30, 2012, from http://tntp.org/assets/misc/20110602_IMPACT_conference_msa_final.pdf?images/uploads/20110602_IMPACT_conference_msa_final.pdf

TNTP. (2012). *Summer report: Creating a culture of excellence in Indiana schools*. Prepared for Indiana Department of Education. Retrieved December 13, 2012, from http://www.riseindiana.org/sites/default/files/files/Summer%20Report.pdf

Tulsa Public Schools. (2011). *TPS Provides 2010–2011 value-added reports to teachers and principals*. Retrieved November 5, 2012, from http://www.tulsaschools.org/2_News/01_PUBLIC_INFO/news_item.asp?ID=13517

Webster, W. J., & Mendro, R. L. (1997). The Dallas value-added accountability system. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 81–99). Thousand Oaks, CA: Corwin Press. Retrieved October 23, 2012, from http://www.dallasisd.org/cms/lib/TX01001475/Centricity/Shared/evalacct/research/articles/Webster-Dallas-Value-added-Accountability-System.pdf

White, S. (2012). *Georgia's SLO process: Lessons learned*. Santa Monica, CA: Teacher Incentive Fund. Retrieved December 13, 2012, from https://www.tifcommunity.org/sites/default/files/Session%205A-%20White-%20Georgia%20SLO%20Process.pdf

Winters, M., Greene, J. P., Ritter, G., & Marsh, R. (2008). *The effect of performance-pay in Little Rock, Arkansas on student achievement*. (Research Brief). Nashville, TN: National Center on Performance Incentives. Retrieved December 14, 2012, from https://my.vanderbilt.edu/performanceincentives/files/2012/10/200802_WintersEtAl_PerfPayLittleRock.pdf