U.S. Department of Education
January 2014

Comparing estimates of teacher value-added based on criterion- and norm-referenced tests

David Stuit
Basis Policy Research

Megan J. Austin
University of Notre Dame

Mark Berends
University of Notre Dame

R. Dean Gerdeman
American Institutes for Research

Key findings

Three analytic strategies were used to compare estimates of teacher value-added based on a criterion-referenced state assessment and a widely used norm-referenced test. They found that:

- Single-year estimates from the state assessment and norm-referenced test were moderately correlated (correlation coefficients of 0.44 to 0.65).
- On average, 33.3 percent of estimates ranked in the same quintile on both tests in the same school year.
- No teachers had estimates above the sample average with 95 percent confidence on one test and below the sample average with 95 percent confidence on the other test.

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

REL MIDWEST
Regional Educational Laboratory
At American Institutes for Research

REL 2014–004

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.

January 2014

This publication was prepared for the Institute of Education Sciences (IES) under contract ED-IES-12-C-0004 by Regional Educational Laboratory Midwest, administered by American Institutes for Research.

Stuit, D., Berends, M., Austin, M. J., & Gerdeman, R. D. (2014). *Comparing estimates of teacher value-added based on criterion- and norm-referenced tests* (REL 2014–004). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midwest. 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

Recent changes to state laws on accountability have prompted school districts to design teacher performance evaluation systems that incorporate student achievement (student growth) as a major component. As a consequence, some states and districts are considering teacher value-added models as part of teacher performance evaluations. Value-added models use statistical techniques to estimate teachers' (or schools') contributions to growth in student achievement over time.

Designers of new performance evaluation systems need to understand the factors that can affect the validity and reliability of value-added results or other measures based on student assessment data used to evaluate teacher performance. This study provides new information on the degree to which value-added estimates of teachers differ by the assessment used to measure their students' achievement growth.

To compare estimates of teacher value-added based on two different assessments, the study selected districts whose students took the criterion-referenced Indiana Statewide Testing for Educational Progress Plus (ISTEP+) and the norm-referenced Measures of Academic Progress (MAP) in the same school year. The analysis examines reading and math achievement data for grades 4 and 5 in 46 schools in 10 Indiana districts for 2005/06–2010/11.

The study used three analytic strategies to quantify the similarities and differences in estimates of teacher value-added from the ISTEP+ and MAP: correlations of value-added estimates based on the two assessments, comparisons of the quintile rankings of value-added estimates on the two assessments, and comparisons of the classifications of value-added estimates on the two assessments according to whether their 95 percent confidence intervals were above, below, or overlapping the sample average.

Consistent with prior research, the study found a moderate relationship between value-added estimates for a single year based on the ISTEP+ and MAP, with average yearly correlation coefficients of 0.44 to 0.65. The comparison of quintile rankings found that an average of 33.3 percent of estimates of teacher value-added ranked in the same quintile on both tests in the same school year. Results were more consistent for estimates in the top and bottom quintiles than in the three middle quintiles. Across all comparisons 28.1 percent of estimates ranked two or more quintiles higher on one test than on the other.

Teacher value-added estimates were more consistent between the ISTEP+ and MAP when considering the precision of the estimates, as measured by confidence intervals. None of the estimates had a 95 percent confidence interval falling above the sample average on one test and a 95 percent confidence interval falling below the sample average on the other.

Overall, the findings indicate variability between the estimates of teacher value-added from two different tests administered to the same students in the same years. Specific sources of the variability across assessments could not be isolated because of limitations in the data and research design. However, the research literature points to measurement error as an important contributor. The findings indicate that incorporating confidence intervals for value-added estimates reduces the likelihood that teachers' performance will be misclassified based on measurement error.

Contents

Why this study?

Improving the evaluation of teacher performance is a top priority for many states and school districts in the Midwest Region and across the country. Since 2009 five of the seven Midwest Region states have developed administrative rules, legislative codes, or state policies requiring districts to implement new systems for evaluating teachers.[1] To comply, school districts must design evaluation systems that include measures of improvement in student achievement (student growth) as a major component (Illinois State Board of Education, 2011a; Wisconsin Department of Public Instruction, 2011). Some districts and states in the region are considering value-added models (Minnesota Department of Education, 2012; Chicago Public Schools, 2011) that use statistical techniques to estimate teachers' (or schools') contribution to growth in student achievement (McCaffrey, Lockwood, Koretz, & Hamilton, 2003).[2]

The findings of this study provide a benchmark for the amount of variability to expect between value-added estimates derived from state test scores and norm-referenced tests

Incorporating value-added or other types of growth measures into evaluations of teacher performance is an intensive process that requires states and districts to build new technological capacities and make important technical decisions (Milanowski, 2011). In addition to designing data management tools that link student and teacher records, a critical step is determining which assessments will be used to measure student growth.

Many districts rely on annually administered state assessments for that purpose. However, districts commonly augment state assessments with commercially available norm-referenced assessments, such as the Northwest Evaluation Association's Measures of Academic Progress (MAP).[3] Norm-referenced tests are attractive for a variety of reasons, including quick turnaround of results, the ability to assess students in grades and subjects not covered by the state assessment, and the opportunity to benchmark student performance against nationally representative peer groups. Further, these tests can be administered multiple times during the school year, so student progress can be monitored at different intervals.

The push to incorporate value-added measures into teacher evaluation systems has prompted Midwest Region states and school districts to ask about the comparability of model results based on different assessments. There is limited empirical evidence on how estimates of teacher value-added vary by assessment: a comprehensive literature review identified only four studies, none conducted in the Midwest Region (Lockwood et al., 2007; Sass, 2008; Corcoran, Jennings, & Beveridge, 2011; Papay, 2011; see appendix A for a detailed review of the literature).

What the study examined

To better understand the implications of using different types of student assessments to evaluate teacher performance, this study provides empirical information on the comparability of estimates of teacher value-added for grades 4 and 5 in Indiana. The estimates are based on the criterion-referenced Indiana Statewide Testing for Educational Progress Plus (ISTEP+) and the norm-referenced MAP, which is widely used in Indiana and other Midwest Region states.

The findings provide a benchmark for the amount of variability to expect between value-added estimates derived from state test scores and norm-referenced tests.[4] To aid in understanding the analyses and interpreting the findings, box 1 defines key terms used in the report, box 2 briefly describes the data and methodology, and appendix B provides more detail on the data and value-added model in this study.

Box 1. Key terms

Confidence interval. A measure of the uncertainty associated with an estimated value, such as an estimate of teacher value-added. A confidence interval indicates the plausible range in which the "true" value lies for a desired level of confidence (for example, 95 percent).

Correlation. The degree to which two measures are related. Correlation coefficients range from –1 to 1, with –1 indicating a perfect negative (inverse) relationship, 1 a perfect positive relationship, and 0 no relationship.

Criterion-referenced test. An assessment designed to measure mastery of a set of content standards and criteria.

Indiana Statewide Testing for Educational Progress Plus (ISTEP+). Indiana's standardized state test, administered annually to all public school students in grades 3–8. It is a criterion-referenced test designed to measure students' mastery of the state's grade-level academic content standards (Indiana Department of Education, 2011).

Measurement error in test scores. The difference between a student's observed test score and the student's "true" ability and knowledge. Several factors affect the degree of measurement error in test scores, including the reliability of the test and the conditions under which the test is administered.

Measures of Academic Progress (MAP). The norm-referenced test used in this study. Districts voluntarily contract with Northwest Evaluation Association, the test developer, to conduct MAP testing.[1] MAP is designed to measure student achievement across a continuum that spans all grades, rather than achievement of specific grade-level content standards. A student's score is usually compared with the scores of a nationally representative peer group in the same grade. MAP is typically administered two to three times per school year (fall and spring; or fall, winter, and spring) to provide teachers and administrators with feedback on students' learning so they can adapt their instruction. MAP tests are "computer-adaptive," meaning that as students take the test on the computer, the test questions adapt based on responses to prior questions. This adaptive process is designed to reliably measure student achievement at all points on the achievement continuum.

Norm-referenced test. An assessment designed to measure how well a student performs relative to other students taking the assessment. A student's score is usually compared with the scores of a nationally representative peer group in the same grade.

Pretest score. The test score used to measure student achievement at the beginning of the school year. It is included in the value-added model to control for baseline achievement when estimating the effect of teacher performance on students' posttest scores.

Posttest score. The test score used to measure student achievement at the end of the school year. It is the outcome measure (dependent variable) used in the value-added model. Estimates of teacher value-added are based on the difference between students' actual posttest scores and their predicted posttest scores from the value-added model.

(continued)

Box 1. Key terms *(continued)*

Quintile. One of five equal-size groups representing one-fifth (20 percent) of a ranked set of data. Estimates of teacher value-added are assigned quintile rankings, from quintile 1, the bottom 20 percent, to quintile 5, the top 20 percent.

Standard deviation. A measure of the amount of variation within a set of data based on how far individual values are from the group average.

Transition matrix. A matrix that plots the proportion of observations within each combination of the row and column categories. This study uses transition matrixes to examine the agreement of quintile rankings of teacher value-added between one test or testing interval and another.

Value-added model. A statistical technique for estimating a teacher's contribution to growth in student achievement by examining changes in test scores over time. This study's value-added model is a covariate adjustment model.

Value-added estimate. An estimate of value-added based on student assessment scores. Because a teacher's true effect (value-added) on student academic outcomes cannot be observed, it is estimated using statistical analysis based on student test scores.

Note

1. The use of MAP is common in the Midwest Region. According to the Northwest Evaluation Association (2011a), more than 1,300 schools in Minnesota administer it. A survey of Wisconsin school districts found that 68 percent conducted periodic testing beyond the required state test, of which 49 percent used MAP (Schug & Niederjohn, 2009). The Ohio Department of Education (2012) approved the use of MAP for providing growth measures for teachers in grades not covered by the state test.

Box 2. Data and methodology

Data. The Indiana Department of Education provided a statewide dataset on students and teachers in grades 4 and 5 in 46 public schools in 10 school districts, including Indiana Statewide Testing for Educational Progress Plus (ISTEP+) scores for 2005/06–2010/11. Scores for the norm-referenced Measures of Academic Progress (MAP) for students in the same 10 school districts for the same period were provided by the test vendor. (The districts already had data-sharing agreements in place with the vendor, Northwest Evaluation Association.) The 10 districts administered the same version of each test in reading and math in both the fall and spring of each school year. The study was thus able to estimate and compare teachers' value-added based on two different assessments administered to the same students in the same school year.

From 2005/06 to 2007/08 the ISTEP+ was administered in the fall, so comparisons of the ISTEP+ and MAP for those years are based on fall-to-fall growth in student achievement. In 2008/09 ISTEP+ testing was switched to the spring, so comparisons in 2009/10 and 2010/11 are based on spring-to-spring growth. The shift in testing date prevented estimating teacher value-added between 2007/08 and 2008/09 because ISTEP+ scores were not available for adjacent spring-to-spring or fall-to-fall testing intervals. Value-added estimates

(continued)

Box 2. Data and methodology *(continued)*

for each teacher for a given academic year are based on the test scores of the same self-contained class of students. For instance, for a grade 4 math teacher's estimate the fall of grade 4 is the pretest and the fall of grade 5 is the posttest. The sample was restricted to grades 4 and 5 because grade 4 is the first grade for which students have prior-year test scores on the ISTEP+, which are necessary to estimate teachers' value-added, and because classrooms for these grades are self-contained, making it easier to attribute students' test scores to the appropriate reading and math instructors.[1] Table B2 in appendix B presents the final samples of teachers and students used in the reading and math analyses.

The data cover 61.6 percent of grade 4 and 5 students in the 10 participating districts over the study period. The other students were excluded because they were missing data elements essential to the analysis (see table B2 in appendix B).[2] The reading sample included 18,787 student observations linked to 1,149 teacher observations over six years. The math sample included 18,787 student observations linked to 1,143 teacher observations. The number of teachers and students varies each year with the availability of ISTEP+ and MAP data and the ability to accurately link student records to their reading and math teachers. The average number of students used to calculate a single-year value-added estimate was 16.4 per teacher in both subjects; the minimum was 11. Teachers matched to 10 or fewer students (3.51 percent of the sample) with sufficient data were excluded because such small samples yield estimates that are imprecise and unlikely to provide reliable information on teacher performance.[3]

Students' ISTEP+ and MAP scaled scores were standardized using the grade- and year-specific means and standard deviations for each subject. Estimates of teacher value-added were adjusted based on indicators of student gender, race/ethnicity, eligibility for free or reduced-price meals, English proficiency status, and special education status. These indicators are included in the value-added model to hold constant student and classroom factors known to influence student achievement but outside a teacher's control.

Methodology. The value-added estimates were calculated using a covariate-adjustment model (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Kane & Staiger, 2008) that predicts the students' posttest scores based on their pretest scores and student and classroom characteristics (see appendix B for details of the model). The average difference between students' predicted and actual posttest scores provides the basis for estimating teacher value-added. Estimates will be positive when students systematically outperform their predicted scores. The ISTEP+ and MAP value-added estimates within each subject and year are based on the same model specification and the same sample of students and teachers. Using the same students to calculate the ISTEP+ and MAP value-added estimates ensures that any differences in estimates are not due to differences in the students that took the two assessments.

Three analytic strategies were used to compare estimates of teacher value-added based on ISTEP+ and MAP scores. One strategy examined correlations of estimates on the two tests. To summarize the results for each subject, the correlations were averaged across years and weighted by the number of teachers in each year. A second strategy constructed transition matrixes to document how the quintile rankings of estimates of teacher value-added differed between the two assessments.[4] For each year, each teacher's estimated value-added was ranked against the estimates for all teachers in the 10 districts for the same grade and subject. Separate transition matrixes were created for fall-to-fall and spring-to-spring estimates

(continued)

Box 2. Data and methodology *(continued)*

in reading and math. A third strategy classified value-added estimates by whether the estimate's 95 percent confidence interval was above, below, or overlapping the average estimate in the sample for the year, subject, and testing interval. This strategy accounts for the statistical uncertainty of the value-added estimates, which is ignored in the transition matrixes and consequently may lead to overstating differences in value-added estimates from the two tests (a teacher can have a large discrepancy in ISTEP+ and MAP quintile rankings even when both estimates are not statistically distinguishable from average).

Notes

1. The four studies most relevant to this project also focused on students in grades 4 and 5 for the same reasons (Koedel & Betts, 2010; Papay, 2011; Lefgren & Sims, 2012; Corcoran et al., 2011).

2. This missing data rate is comparable to that for other studies that link longitudinal student and teacher records (Corcoran et al., 2011; Papay, 2011; Lockwood et al., 2007). School districts suffer the same missing data problems in their value-added modeling (Papay, 2011).

3. This restriction is consistent with the value-added research literature and approaches used by school districts (Papay, 2011; Corcoran et al., 2011; Isenberg & Hock, 2011; Sass, Hannaway, Xu, Figlio, & Feng, 2012; Value-Added Research Center, 2010).

4. Transition matrixes based on quintile or quartile rankings are commonly used to study the stability of estimates of teacher value-added (Koedel & Betts, 2007; Goldhaber & Hansen, 2010; McCaffrey, Sass, Lockwood, & Mihaly, 2009; Ballou, 2005; Aaronson, Barrow, & Sander, 2007).

Overall, the findings indicate yearly variability between the two tests in estimates of teacher value-added

Findings

Overall, the findings indicate yearly variability between the two tests in estimates of teacher value-added. Consistent with prior research, the study found a moderate relationship between value-added estimates for a single year based on the ISTEP+ and MAP, with average yearly correlation coefficients of 0.44 to 0.65 (table 1). The comparison of quintile rankings found that an average of 33.3 percent of estimates of teacher value-added ranked in the same quintile on both tests in the same school year. However, across all comparisons, 28.1 percent of estimates had at least a two-quintile difference between ISTEP+ and MAP scores. Teacher performance classifications were more consistent for confidence intervals: none of the estimates had a 95 percent confidence interval above the sample average on one test and a 95 percent confidence interval below the sample average on the other.

Correlations of value-added estimates on the two tests are moderate

The correlations of estimates of teacher value-added based on the ISTEP+ and MAP are moderate to moderately strong by conventional standards (Cohen, 1988; Hemphill, 2003). Average yearly correlations ranged from 0.35 (grade 4 fall to fall) to 0.64 (grade 5 spring to spring) for reading and from 0.43 (grade 4 fall to fall) to 0.72 (grade 5 spring to spring) for math (see table 1). These correlations are consistent in magnitude with those reported in three similar studies that compared teachers' value-added estimates from different assessments (Sass, 2008; Papay, 2011; Corcoran et al., 2011).[5]

Average correlations are higher in math than in reading, which is also consistent with prior research (McCaffrey et al., 2009; Corcoran et al., 2011; Sass et al., 2012). On both tests the standard deviations of value-added estimates are larger in math than in reading, indicating that value-added estimates are more dispersed from the sample average in math

Table 1. Correlations of estimates of teacher value-added based on the ISTEP+ and MAP, 2005/06–2010/11

Year/grade	Reading		Math	
	Fall-to-fall ISTEP+ and MAP	Spring-to-spring ISTEP+ and MAP	Fall-to-fall ISTEP+ and MAP	Spring-to-spring ISTEP+ and MAP
Year				
2005/06	0.37	na	0.50	na
2006/07	0.57	na	0.54	na
2007/08	—	na	—	na
2008/09	na	—	na	—
2009/10	na	0.51	na	0.60
2010/11	na	0.61	na	0.67
Grade				
4	0.35	0.49	0.43	0.57
5	0.55	0.64	0.61	0.72
Average	0.44	0.56	0.52	0.65

ISTEP+ is the Indiana Statewide Testing for Educational Progress Plus. MAP is the Measures of Academic Progress assessment that was administered in the same schools during the same period as the ISTEP+.

na is not applicable because tests were administered during a different interval that year.

— is not available because in 2008/09 the administration of the ISTEP+ switched from the fall to the spring, so ISTEP+ scores were not available for adjacent spring-to-spring or fall-to-fall testing intervals.

Note: Values are Pearson correlation coefficients (r). Each teacher's set of value-added estimates for a given academic year is based on the test scores of the same self-contained class of students. For instance, for a grade 4 teacher's estimate, the fall of grade 4 is the pretest and the fall of grade 5 is the posttest.

Source: Authors' calculations based on data from the Indiana Department of Education and Northwest Evaluation Association.

than in reading (see table C2 in appendix C). For example, the average standard deviation of spring-to-spring MAP value-added estimates is 0.204 in math and 0.148 in reading.

In both subjects correlations are higher for the spring-to-spring testing interval than for the fall-to-fall interval. This study does not provide evidence on why this pattern is observed. Papay (2011), who also found lower correlations in fall-to-fall comparisons, attributed that to weaker reliability in students' fall test scores, perhaps due to the effects of summer learning loss.

A third of value-added estimates rank in the same quintile on both tests in the same school year

Across both subjects and testing intervals, one-third of quintile rankings of estimates of teacher value-added based on the ISTEP+ and MAP were identical (table 2). An additional 38.5 percent of rankings were within one quintile of each other. Some 18.9 percent of estimates differed by two quintiles, indicating that students performed substantially better on one test than on the other.[6] An additional 7.2 percent differed by three quintiles, and 2.0 percent ranked in the top quintile on one assessment and the bottom quintile on the other.

Estimates of teacher value-added in the top and bottom quintiles are more likely to maintain their ranking on both tests than estimates in quintiles 2–4 (table 3). For example, the spring-to-spring matrixes indicate that 49.5 percent of value-added estimates ranked in the

Table 2. Summary of agreement between quintile rankings of estimates of teacher value-added based on the ISTEP+ and MAP, 2005/06–2010/11 (percent)

Subject and testing period	Same quintile ranking	One-quintile difference	Two-quintile difference	Three-quintile difference	Four-quintile difference
Reading					
Fall-to-fall ISTEP+ and MAP	26.3	36.2	22.7	10.3	4.5
Spring-to-spring ISTEP+ and MAP	35.6	37.7	19.3	6.6	0.8
Math					
Fall-to-fall ISTEP+ and MAP	33.8	39.8	17.9	7.4	1.2
Spring-to-spring ISTEP+ and MAP	37.6	40.6	15.9	4.4	1.5
Average	33.3	38.5	18.9	7.2	2.0

ISTEP+ is the Indiana Statewide Testing for Educational Progress Plus. MAP is the Measures of Academic Progress that was administered in the same schools to the same students and during the same period as the ISTEP+.

Note: Values may not sum to 100 percent because of rounding.

Source: Authors' calculations based on data from the Indiana Department of Education and Northwest Evaluation Association.

Table 3. Agreement of quintile rankings of estimates of teacher value-added based on the ISTEP+ and MAP, 2005/06–2010/11 (percent)

Reading

		Fall-to-fall MAP				
		Bottom	2nd	3rd	4th	Top
Fall-to-fall ISTEP+	Bottom	**36.2**	27.5	14.3	12.3	9.7
	2nd	21.9	**17.6**	26.2	24.5	9.8
	3rd	11.6	25.6	**20.2**	16.9	25.6
	4th	17.4	17.1	19.1	**24.5**	22.0
	Top	12.8	12.2	20.2	21.8	**32.9**

		Spring-to-spring MAP				
		Bottom	2nd	3rd	4th	Top
Spring-to-spring ISTEP+	Bottom	**52.0**	25.1	16.7	3.1	3.2
	2nd	20.0	**25.9**	24.0	18.6	11.6
	3rd	18.0	21.7	**24.0**	23.7	12.6
	4th	9.0	18.1	22.9	**26.8**	23.2
	Top	1.0	9.2	12.5	27.8	**49.5**

Math

		Fall-to-fall MAP				
		Bottom	2nd	3rd	4th	Top
Fall-to-fall ISTEP+	Bottom	**47.1**	24.4	14.5	11.6	2.4
	2nd	27.1	**30.5**	18.1	13.4	11.0
	3rd	16.5	23.2	**30.1**	16.8	13.4
	4th	5.9	13.4	19.3	**24.8**	36.6
	Top	3.5	8.5	18.1	33.3	**36.6**

		Spring-to-spring MAP				
		Bottom	2nd	3rd	4th	Top
Spring-to-spring ISTEP+	Bottom	**47.9**	28.7	16.0	4.2	3.2
	2nd	31.3	**28.4**	18.1	13.7	8.6
	3rd	13.5	22.8	**28.7**	25.3	9.7
	4th	3.1	13.9	24.5	**31.6**	26.9
	Top	4.2	6.2	12.8	25.3	**51.6**

ISTEP+ is the Indiana Statewide Testing for Educational Progress Plus. MAP is the Measures of Academic Progress assessment that was administered in the same schools and during the same period as the ISTEP+.

Note: Values are the percentage of estimates of teacher value-added that fall into the quintile rankings indicated by the corresponding row and column headings. Within each matrix, the values in each column and row sum to 100 percent. The bolded diagonal cells indicate the percentages of value-added estimates that fall into the same quintile for both assessments. Perfect alignment of estimates would be indicated by values of 100 in each diagonal cell, indicating that 100 percent of estimates fall into the same quintile on each assessment and testing interval combination.

Source: Authors' calculations based on data from the Indiana Department of Education and Northwest Evaluation Association.

top quintile on both MAP and ISTEP+ reading tests and 51.6 percent did on both MAP and ISTEP+ math tests. These findings align with other research evidence indicating that value-added estimates are most reliable when used to distinguish the highest and lowest performing teachers (Milanowski, Heneman, & Kimball, 2011). Goldhaber and Hansen (2010) find that high-performing teachers have more stable value-added estimates from year to year.

The most consistent classifications of value-added estimates were based on statistical confidence

The majority of yearly estimates of teacher value-added were not distinguishable from the sample average with 95 percent confidence (table 4). In reading, 96.4 percent of the fall-to-fall estimates and 70.9 percent of spring-to-spring estimates were classified as average, meaning the estimates were not statistically distinguishable from average on either test. Differentiation was greater for math, with 65.0 percent of fall-to-fall estimates and 47.2 percent of spring-to-spring estimates classified as average on both tests.

The majority of yearly estimates of teacher value-added were not distinguishable from the sample average with 95 percent confidence

ISTEP+ and MAP value-added estimates were classified in the same range (average, above average, or below average) for the majority of teachers in the sample (table 5). Across both subjects 76.2 percent of all pairs of yearly value-added estimates had identical classifications, with 69.2 percent classified as average on both tests, 3.7 percent classified as below average on both tests, and 3.3 percent classified as above average on both tests. No teacher yearly value-added estimates were classified as above average on one test and below average on the other.

Implications of the findings

This research provides new evidence on the comparability of estimates of teacher value-added based on different tests. Overall, the study finds a moderate to moderately strong relationship between teachers' yearly value-added from the criterion-referenced ISTEP+ and the norm-referenced MAP, with average correlations of 0.44 to 0.65. These correlations are consistent with those reported in three previous studies that compared value-added estimates from different tests using similar methods (Sass, 2008; Papay, 2011; Corcoran et al., 2011). Correlations between other measures of teacher performance may be lower than the correlations between value-added estimates found in this study. For example, Ho and Kane (2013) reported correlations ranging from 0.37 to 0.47[7] for teacher classroom observation scores from different types of raters, and Hill et al. (2012) reported correlations of 0.26 to 0.44 between classroom observation ratings and teacher scores on a pedagogical content knowledge assessment.

Due to limitations in the data and research design (see below), the precise sources of the variability in value-added estimates across the two tests cannot be identified. However, it is possible to rule out some potential sources. Because the same groups of students are used to estimate teacher value-added for both tests, differences in student and classroom characteristics cannot be a source of variability. Further, a supplemental analysis indicates that differences in the test content of the ISTEP+ and MAP are unlikely to be a major contributing factor because individual students' test scores on the ISTEP+ and MAP are highly correlated (see appendix B for details).

The research literature points to measurement error as an important source of the variability in estimates of teacher value-added. Measurement errors arise from factors that affect test-taking conditions for a particular student, class, or school on a given day, such as the

Table 4. Estimates of teacher value-added in the average, above average, and below average ranges, based on the ISTEP+ and MAP, 2005/06–2010/11 (percent)

Reading

Fall-to-fall ISTEP+ Range	Fall-to-fall MAP: Above average	Average	Below average	Row total
Above average	**0.0**	1.2	0.0	1.2
Average	1.0	**96.4**	0.5	97.8
Below average	0.0	1.0	**0.0**	1.0
Column total	1.0	98.6	0.5	

Spring-to-spring ISTEP+ Range	Spring-to-spring MAP: Above average	Average	Below average	Row total
Above average	**3.0**	7.2	0.0	10.1
Average	3.9	**70.9**	5.7	80.5
Below average	0.0	7.1	**2.2**	9.4
Column total	6.9	85.3	7.9	

Math

Fall-to-fall ISTEP+ Range	Fall-to-fall MAP: Above average	Average	Below average	Row total
Above average	**1.7**	14.3	0.0	15.9
Average	1.9	**65.0**	3.4	70.3
Below average	0.0	10.9	**2.9**	13.8
Column total	3.6	90.1	6.3	

Spring-to-spring ISTEP+ Range	Spring-to-spring MAP: Above average	Average	Below average	Row total
Above average	**8.0**	9.6	0.0	17.6
Average	8.3	**47.2**	8.5	64.1
Below average	0.0	9.0	**9.3**	18.3
Column total	16.3	65.8	17.8	

ISTEP+ is the Indiana Statewide Testing for Educational Progress Plus. MAP is the Measures of Academic Progress assessment. MAP was administered in the same schools and classrooms and during the same period as the ISTEP+.

Note: Values may not sum to 100 percent because of rounding. Estimates are classified as above average if the lower bound of their 95 percent confidence interval is greater than the average value-added estimate in the sample of that particular year, subject, and testing interval; estimates are classified as average if their 95 percent confidence interval overlaps the sample average; estimates are classified as below average if the upper bound of their 95 percent confidence interval is below the sample average. The bolded diagonal cells indicate the percentages of value-added estimates that fall into the same range for both assessments. Perfect alignment of estimates would be indicated by values of 100 in each diagonal cell, indicating that 100 percent of estimates fall into the same range on each assessment and testing interval combination. A minimum of 11 students per class is necessary to include the teacher in the analysis; teachers matched to 10 or fewer students (3.51 percent of the sample) with sufficient data were excluded because estimates of teacher value-added estimates derived from such small samples are too imprecise.

Source: Authors' calculations based on data from the Indiana Department of Education and Northwest Evaluation Association.

Table 5. Summary of agreement between estimates of teacher value-added in the average, above average, and below average ranges, based on the ISTEP+ and MAP, 2005/06–2010/11 (percent)

Subject and testing period	Average on both tests	Below average on both tests	Above average on both tests	Overall agreement rate
Reading				
Fall-to-fall ISTEP+ and MAP	96.4	0.0	0.0	96.4
Spring-to-spring ISTEP+ and MAP	70.9	2.2	3.0	76.1
Average	83.7	1.1	1.5	86.3
Math				
Fall-to-fall ISTEP+ and MAP	65.0	2.9	1.7	69.6
Spring-to-spring ISTEP+ and MAP	47.2	9.3	8.0	64.5
Average	56.1	6.1	4.9	67.1
Overall average	69.2	3.7	3.3	76.2

ISTEP+ is the Indiana Statewide Testing for Educational Progress Plus. MAP is the Measures of Academic Progress that was administered in the same schools to the same students and during the same period as the ISTEP+.

Note: Values indicate the percentage of pairs of ISTEP+ and MAP yearly value-added estimates that were classified in the same range (average, below average, above average) on both tests.

Source: Authors' calculations based on data from the Indiana Department of Education and Northwest Evaluation Association.

time of day the test is administered or the behavior of classmates during the test (Goldhaber & Hansen, 2010; Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2008; Corcoran et al., 2011). A frequently cited example is a loudly barking dog in the school parking lot on test day (Kane & Staiger, 2008). Such idiosyncratic factors as these will cause student test scores, and consequently estimates of teacher value-added, to fluctuate from year to year and test to test for reasons unrelated to the quality of teaching and learning in the classroom.

Policymakers need to consider measurement error when interpreting estimates of teacher value-added, particularly single-year estimates. This study finds that incorporating confidence intervals into teacher performance classifications reduces the likelihood that teachers classified as higher performing based on one test will be classified as lower performing based on another. Other research has shown that averaging estimates of teacher value-added over multiple years will increase precision and provide more reliable information on teacher performance (McCaffrey et al., 2009).

Incorporating confidence intervals into teacher performance classifications reduces the likelihood that teachers classified as higher performing based on one test will be classified as lower performing based on another

Limitations

This study has several limitations, especially when it comes to state and district education policymakers' use of empirical studies in decisionmaking.

Missing data may lead to estimates that do not reflect the teachers and students in the study sample. In practice, states and school districts face similar challenges with missing data in value-added modeling (Papay, 2011). The sample for this study included 61.6 percent of the students in grades 4 and 5 across 46 schools in 10 participating Indiana districts for 2005/06–2010/11. The missing data rate is comparable to that in other studies that link longitudinal data on students and their teachers (Corcoran et al., 2011; Papay, 2011; Lockwood et al., 2007).

The characteristics of the sample limit the generalizability of the research findings. The analysis was limited to teachers and students in grades 4 and 5, so results may not generalize to other grade levels. For example, in middle school (grades 6–8), students are often in classes taught by subject-specific teachers for each semester of the school year or take specialized classes within the same subject at the same time (for example, pre-algebra and geometry). The data available for this study do not include the information required to estimate value-added for middle school teachers.

The findings also may not generalize to other districts in Indiana or other states. The analysis was based on data from 10 of the state's 347 districts. The 46 schools in the sample enrolled proportionally fewer economically disadvantaged students than the average, and average ISTEP+ test scores were substantially above the state average (see table B3 in appendix B). Expanding the analyses to other districts in Indiana or in other states would strengthen the generalizability of the findings.

Because of data limitations, the analysis was based on a comparison of single-year value-added estimates. A limitation of using single-year estimates is that it is not possible to distinguish the persistent effects of teachers on student achievement over time from the factors that are specific to their classrooms in each year. While the value-added model controls for some measurable classroom characteristics, such as average prior test scores,

these may not account for all the classroom-specific factors that influence student achievement. The study findings may not generalize to teachers' long-term value-added estimates from different assessments.

Finally, the study could not determine the sources of the variability in value-added estimates between tests. Differences in the measurement characteristics of the ISTEP+ and MAP are one potential source. State tests such as the ISTEP+ are designed to assess whether students are proficient in the state standards, and they are calibrated to maximize the precision of scores that are close to the state proficiency standard. Therefore, these tests typically exhibit "ceiling" and "floor" effects that limit their ability to differentiate the performance of students at the very high or low ends of the achievement spectrum. MAP, as is characteristic of norm-referenced tests, is designed to measure all points across the achievement range with similar precision. Another potential source of variability is differences in effort that students and teachers exert on the ISTEP+ and MAP, considering that the ISTEP+ is a high-stakes test used for school accountability purposes, such as adequate yearly progress ratings. Investigating these hypotheses was outside the scope of this study.

Appendix A. Literature review

Four studies were identified that examined the ways estimates of teacher value-added differ across tests (Lockwood et al., 2007; Sass, 2008; Corcoran et al., 2011; Papay, 2011). Collectively, these studies indicate moderate correlation between value-added estimates for different tests. However, some of the correlations were found to be affected by factors such as the stakes attached to tests, test timing, and measurement error. For example, using the same sample of students and up to eight years of data for each teacher, one study of value-added estimates for reading and math on a high-stakes test and a low-stakes test found moderate correlations, but the teacher effects were 15–31 percent larger on the high-stakes tests (Corcoran et al., 2011).

Another study compared student test scores on three reading achievement tests: a state reading assessment, the Stanford 9 Achievement Test, and the Scholastic Reading Inventory (Papay, 2011). Correlations among the tests ranged from 0.15 to 0.58, indicating that teachers who promoted achievement growth on one measure also promoted growth on the others. Despite these generally small to moderate correlations, based on the conventional standards of Cohen (1988) and Hemphill (2003), test timing and measurement error explained more of the differences in value-added estimates than did test content, score scaling, or differences among students in each sample.

Papay (2011) also investigated the stability of value-added estimates related to the timing of test administrations and found variations in estimates based on when the pretests and posttests were administered. For example, the correlation of value-added estimates between the Stanford 9 Achievement Test and the Scholastic Reading Inventory was greater when measured from one fall to the next fall (0.27) than from fall to spring (0.12). For the same test (Scholastic Reading Inventory) the rank correlation was 0.21 between fall to spring and fall to fall and –0.06 between spring to spring and fall to fall. According to Papay, these comparisons suggest the importance of taking summer learning loss into account when estimating teacher effectiveness.

Appendix B. About the data and the value-added model

This appendix describes the data and the value-added model used in the study.

About the data

Data sources. The Indiana Department of Education provided the statewide dataset of public school students' scores on the Indiana Statewide Testing for Educational Progress Plus (ISTEP+) for 2005/06–2010/11. The dataset contains student-level demographic characteristics, including gender, race/ethnicity, and three binary indicators (eligibility for free or reduced-price meals, English proficiency status, and special education status). Students eligible for the federal free or reduced-price meals program are from households that meet the income guidelines (household income at no more than 130 percent of the poverty line for free lunch and no more than 180 percent for reduced-price lunch). English proficiency status identifies students who are not fluent in English as determined by the Links English Language Proficiency Assessment. Special education status identifies students who receive special education services. These variables are used as controls in estimating teacher value-added. Northwest Evaluation Association, the vendor of the norm-referenced Measures of Academic Progress assessment (MAP), provided scores for 2005/06–2010/11 for students in select Indiana school districts with which it had preexisting data-sharing agreements.

The Indiana Department of Education and the Northwest Evaluation Association provided class roster data with unique identifiers so that students could be matched to their reading and math teachers while their confidentiality remained protected. School districts reported the student-teacher links in the class rosters to the department and the vendor, so the accuracy of the links depends on the districts' reporting protocols. Teachers were designated as students' reading or math instructors if they were listed as the teacher of record in the class roster files for reading or math courses or if they were recorded as the sole teacher of all general education courses in a self-contained classroom. Students with more than one teacher recorded for a reading or math course and students with team teachers in self-contained classrooms were dropped from the sample.

The sample. The sample included teachers and students in 46 schools in 10 Indiana school districts (table B1). To compare estimates of teacher value-added based on results for the two tests for the same students in the same years, districts were selected if they administered both the ISTEP+ and MAP in the same academic years. The analysis was limited to students and teachers in grades 4 and 5. Most of the studies relevant to this project also focused on students in grades 4 and 5, including Koedel and Betts (2010), Papay (2011), Lefgren and Sims (2012), and Corcoran et al. (2011).

There are two reasons for restricting the sample to grades 4 and 5. First, most students in these grades were in self-contained classrooms, which makes it easier to match their test score records to the appropriate reading and math teachers. By contrast, most middle school students (grades 6–8) in these districts were in departmentalized classes and often switched subject teachers over semesters of the same school year. These students were also enrolled in multiple specialized courses within the same subject (such as pre-algebra and geometry) in the same school year. The course roster data did not have sufficient detail to overcome these complexities and accurately determine which middle school students' reading and math scores should be attributed to a given teacher. Second, students in

Table B1. Number of student and teacher observations included in the analysis, by grade, 2005/06–2010/11

Year and grade	Reading			Math		
	Number of student observations	Number of teacher observations	Average number of students per teacher[a]	Number of student observations	Number of teacher observations	Average number of students per teacher[a]
Year						
2005/06	4,113	263	15.6	4,142	263	15.7
2006/07	2,398	156	15.4	2,387	151	15.8
2007/08	2,114	134	15.8	2,243	141	15.9
2008/09	2,950	190	15.5	2,947	190	15.5
2009/10	3,434	220	15.6	3,300	209	15.8
2010/11	3,778	186	20.3	3,768	189	19.9
Grade						
4	10,296	629	16.4	10,271	628	16.4
5	8,491	520	16.3	8,516	515	16.5
Total	18,787	1,149	16.4	18,787	1,143	16.4

Note: Based on 46 schools in 10 Indiana public school districts.

a. Based on the number of students used to estimate teacher value-added.

Source: Authors' analysis based on data from the Indiana Department of Education and Northwest Evaluation Association.

grades 4 and 5 had pre- and posttest scores on both the ISTEP+ and MAP in the same semesters (fall or spring), which enabled comparisons of the value-added estimates. ISTEP+ testing starts in grade 3, so grade 3 students were excluded from the sample as they would not have spring pretest scores for grade 2; however, spring-to-spring estimates for grade 4 students use grade 3 test scores as pretests.

Overall, the sample included 61.6 percent of the students in grades 4 and 5 enrolled in the 10 participating districts from 2005/06 to 2010/11. The missing data rate of 38.4 percent is comparable to that in other studies that require linking longitudinal student and teacher records (Corcoran et al., 2011; Papay, 2011; Lockwood et al., 2007). The percentage of grade 4 and 5 students dropped from the sample because it was not possible to match their ISTEP+ scores to their MAP scores in the same subject within the same year and test administration time (spring or fall) was 16.9 percent for reading and 17.1 percent for math (table B2). The percentage dropped from the sample because it was not possible to match students' scores from the same test across consecutive school years in order to establish pre-and posttest scores was 13.6 percent for reading and 13.5 percent for math. And the percentage dropped because students could not be matched to classroom teachers was 7.9 percent for reading and 7.8 percent of for math.

Average student characteristics for the 46 schools represented in the study differed from the average for all public schools in Indiana that served grades 4 and 5 from 2005/06 to 2010/11 (table B3). On average, the students enrolled in the 46 schools had higher achievement levels than their peers across the state, with the median percentile rank of ISTEP+ scores exceeding the state average in all grades and subjects. The schools included in the study had an average percentage of students eligible for free or reduced-price meals below

Table B2. Number of student observations contributing to teacher value-added estimates, by subject, 2005/06–2010/11

Subject and category	2005/06	2006/07	2007/08	2008/09	2009/10	2010/11	Total	Cumulative share of data dropped (percent)
Reading								
Total number of enrolled students (grades 4 and 5)	5,324	5,310	4,843	4,905	5,077	5,023	30,482	
After matching ISTEP+ and MAP records within the same school year	4,552	4,248	3,826	4,101	4,275	4,315	25,317	16.9
After matching ISTEP+ and MAP pre- and posttest scores from adjacent years	4,229	3,254	2,869	3,317	3,587	3,926	21,182	30.5
After student–teacher linking	4,113	2,398	2,114	2,950	3,434	3,778	18,787	38.4
Share of total enrolled students included in analysis (percent)	77.3	45.2	43.7	60.1	67.6	75.2	61.6	38.4
Math								
Total number of enrolled students (grades 4 and 5)	5,324	5,310	4,843	4,905	5,077	5,023	30,482	
After matching ISTEP+ and MAP records within the same school year	4,579	4,168	3,826	4,120	4,265	4,320	25,278	17.1
After matching ISTEP+ and MAP pre- and posttest scores from adjacent years	4,258	3,168	2,869	3,337	3,582	3,931	21,145	30.6
After student–teacher linking	4,142	2,387	2,243	2,947	3,300	3,768	18,787	38.4
Share of total enrolled students included in analysis (percent)	77.8	45.0	46.3	60.1	65.0	75.0	61.6	38.4

ISTEP+ is the Indiana Statewide Testing for Educational Progress Plus. MAP is the Measures of Academic Progress assessment that was administered in the same schools and during the same period as the ISTEP+.

Source: Authors' calculations based on data from the Indiana Department of Education and Northwest Evaluation Association.

the state average and enrolled proportionally fewer students receiving special education services, students with limited English proficiency, and students of a racial/ethnic minority.

About the ISTEP+ and MAP. The ISTEP+ and MAP test scores used in this study are scaled scores. The ISTEP+ is administered annually to all Indiana public school students in grades 3–8. It is a criterion-referenced test designed to provide a summative assessment of students' mastery of the state's grade-specific academic content standards (Indiana Department of Education, 2011). The ISTEP+ is vertically equated across grades and consists of multiple-choice, constructed-response, and extended-response items scored using item-response theory methods (Indiana Department of Education, 2011). Reliability coefficients ranged from 0.88 to 0.94 for reading and 0.88 to 0.95 for math (Indiana Department of Education, 2011).

MAP is also vertically scaled. For each subject the test scores of all grade levels are placed on a single scale developed using item-response theory (Kingsbury, 2003; Northwest Evaluation Association, 2011b). MAP is a computer-adaptive test that consists of 40–60 items. Test items adapt in difficulty depending on student performance on particular items. Reliability coefficients for MAP tests range from 0.94 to 0.95 in reading and from 0.92 to 0.97 in math (Cronin, 2005). All ISTEP+ and MAP scaled scores are standardized to a mean of zero and a standard deviation of one using the grade- and year-specific means and standard deviations of each assessment subject.

Table B3. Average student characteristics of schools included in the analysis compared with state averages, 2005/06–2010/11

Characteristic	Schools included in analysis	State average[a]
Receiving free or reduced-price meals (percent)	38.8	52.8
With limited English proficiency (percent)	2.7	5.2
Receiving special education services (percent)	11.4	14.2
Black (percent)	6.3	12.1
Hispanic (percent)	3.4	8.5
Female (percent)	49.8	49.1
Median state percentile rank on ISTEP+[b]		
Reading, grade 4, spring	57	48
Reading, grade 5, spring	56	48
Math, grade 4, spring	58	49
Math, grade 5, spring	57	48
Average enrollment, grades 4 and 5[c]	90.8	133.6
Number of schools	46	1,142
Number of student observations	18,787	935,488

ISTEP+ is the Indiana Statewide Testing for Educational Progress Plus.

Note: Values are school-level averages of the characteristics of all students in grades 4 and 5 with ISTEP+ test scores from 2005/06 to 2010/11.

a. Based on all public schools serving students in grades 4 and 5, including public charter schools.

b. Based on students' median percentile rank of students' ISTEP+ scores relative to all other Indiana public school students in the same grade, year, and subject. It indicates the percentile rank of the student at the 50th percentile in the school at the 50th percentile within the sample.

c. Based on the number of students in grades 4 and 5 observed in the ISTEP+ data received from the Indiana Department of Education.

Source: Authors' calculations based on data from the Indiana Department of Education and Northwest Evaluation Association.

ISTEP+ testing dates are determined by the state, whereas MAP dates are at the discretion of the school districts. From 2005/06 to 2007/08 the ISTEP+ was administered during the last two weeks of September. In 2008/09 ISTEP+ testing was switched from the fall to the last week of April and first week of May. The 10 districts in the study administered MAP in both the fall and spring of each school year. With one exception, all districts' fall and spring testing windows for MAP are within one week of the ISTEP+ fall and spring testing windows (table B4).[8]

The value-added model

A covariate-adjustment model was used to estimate teacher value-added (McCaffrey et al., 2004). At the time of this study, covariate adjustment models were in use in large school districts in the Midwest and elsewhere, including Chicago Public Schools, the Madison Metropolitan School District, and Milwaukee Public Schools (Chicago Public Schools, 2012; Value-Added Research Center, 2012a; Value-Added Research Center, 2012b). The District of Columbia Public Schools' IMPACT evaluation system and Florida's state evaluation model use variants of the covariate-adjustment model.[9]

The model specifies the posttest score as a linear function of the pretest score and individual and classroom-level characteristics. Separate models are fit for each year, grade, and subject. Models are run separately for each test (ISTEP+ and MAP) and testing interval

Table B4. ISTEP+ and MAP testing windows, by school district, 2005/06–2010/11

	Fall testing window		Spring testing window	
District	ISTEP+	MAP	ISTEP+	MAP
A	Last two weeks of September	First two weeks of September	Last week of April and first week of May	Second and third weeks of April
B	Last two weeks of September	Second and third weeks of September	Last week of April and first week of May	Second and third weeks of April
C	Last two weeks of September	First two weeks of September	Last week of April and first week of May	Last week of April
D	Last two weeks of September	Last week of August	Last week of April and first week of May	Last week of April and first week of May
E	Last two weeks of September	Second and third weeks of September	Last week of April and first week of May	First week of May
F	Last two weeks of September	Last week of September and first week of October	Last week of April and first week of May	Second and third weeks of April
G	Last two weeks of September	First two weeks of September	Last week of April and first week of May	Second and third weeks of April
H	Last two weeks of September	First two weeks of September	Last week of April and first week of May	Last week of April and first week of May
I	Last two weeks of September	Second and third weeks of September	Last week of April and first week of May	Last week of April and first week of May
J	Last two weeks of September	Second and third weeks of September	Last week of April and first week of May	Last week of April and first week of May
Years	2005/06–2007/08	2005/06–2010/11	2008/09–2010/11	2005/06–2010/11

ISTEP+ is the Indiana Statewide Testing for Educational Progress Plus. MAP is the Measures of Academic Progress assessment that was administered in the same schools and during the same period as the ISTEP+.

Source: Indiana Department of Education and Northwest Evaluation Association.

(fall to fall and spring to spring) but use the same samples of students and teachers within the grade, year, and subject combination.

Teacher observations cannot be linked across school years, so teachers' persistent value-added effects could not be calculated across multiple years.[10] All estimates of single-year teacher value-added are based on the following model specification:

$$Y_{ij}^{t} = \lambda \mathbf{Y}_{ij}^{t-1} + \beta \mathbf{X}_{ij}^{t} + \zeta \overline{\mathbf{X}}_{j}^{t} + \theta_{j}^{t} + \varepsilon_{ij}$$

where Y_{ij}^{t} is the math or reading posttest score in year t for student i with teacher j; $\mathbf{Y}_{i}^{t-1}$ is a vector of pretest scores that includes the student's pretest score in the same subject as the posttest and the pretest score in the off subject (for example, the reading score is the off-subject pretest score when the math score is the outcome); $\mathbf{X}_{i}$ is a vector of student-level characteristics that includes binary indicators of gender, race/ethnicity (Black, Hispanic), eligibility for free or reduced-price meals, English proficiency status, and special education status; $\overline{\mathbf{X}}_{j}$ is a vector of the classroom-level means of the student characteristics and pretest scores, which are included to control for the effects of classroom composition on student achievement (for example, a high concentration of limited English proficient students);[11] θ_{j}^{t} is a vector of teacher random effects in year t; and ε is a student-specific random error term.

This study is interested in the teacher effects, θ_{j}, which are assumed to be independent and identically distributed normal random variables with zero means.[12] A hierarchical linear

model is used to compute the best linear unbiased predictions of the random effects (Robinson, 1991; Raudenbush & Bryk, 2002). The hierarchical linear model shrinks estimates of teacher value-added back to the sample mean according to their reliability, which is a function of the number of student test score observations and the error variance of the estimate (Raudenbush & Bryk, 2002; McCaffrey et al., 2009). Estimates with lower reliability (fewer students and higher variance) are pulled closer to the mean than are estimates with higher reliability.

Appendix C. Supplemental analysis of correlations of students' scores on the Indiana Statewide Testing for Educational Progress Plus and Measures of Academic Progress

Correlations of students' posttest scores in the same subject and from the same test administration time were examined to investigate sources of discrepancies between estimates of teacher value-added based on scores on the criterion-referenced Indiana Statewide Testing for Educational Progress Plus (ISTEP+) and the norm-referenced Measures of Academic Progress assessment (MAP). The analysis provides evidence on the alignment of the academic content of the ISTEP+ and MAP, which helps in discerning the extent to which differences in value-added estimates for the two tests stem from differences in test content rather than from other factors such as measurement error. If the ISTEP+ and MAP measure different academic competencies, the discrepancies between estimates could reflect differences in teachers' instructional focus during the school year.

The correlations of the ISTEP+ and MAP scores suffer from attenuation bias due to measurement error inherent in all test scores. To disattenuate the correlations of measurement error, the correlations were divided by the square root of the product of the ISTEP+ and MAP reliability coefficients (Spearman, 1904). This is done using the total test internal consistency reliability coefficients (Cronbach alphas) from the ISTEP+, which fall between 0.90 and 0.92 in both subjects for grades 4 and 5 (Indiana Department of Education, 2011). For MAP reliability estimates the test publisher's marginal reliability coefficients were used; they provide internal consistency measures for non-fixed-form tests and range from 0.93 to 0.97 for reading and math in grades 4 and 5.

Even after disattenuation, the correlations may understate the content alignment between the two tests because of idiosyncratic factors that influence student performance on the two tests (Papay, 2011). For example, a student might get a better night's sleep before one test than before the other.

To gather evidence on the alignment of the test content of the ISTEP+ and MAP, an additional test—proposed by Papay (2011)—was conducted that compared correlations of students' scores on the same test over time to correlations of their scores on different tests over time. If the two tests were measuring different content, the year-to-year correlations of students' scores on the same test would be expected to exceed the year-to-year correlations of their scores on different tests. For example, the correlation of students' ISTEP+ scores at time t and time t+1 are expected to exceed the correlations between their ISTEP+ scores at time t and their MAP scores at time t+1. Formally, this test is:

$$\text{Corr}(\text{ISTEP+}_{t}, \text{ISTEP+}_{t+1}) > \text{Corr}(\text{ISTEP+}_{t}, \text{MAP}_{t+1})$$

$$\text{Corr}(\text{MAP}_{t}, \text{MAP}_{t+1}) > \text{Corr}(\text{MAP}_{t}, \text{ISTEP}_{t+1}).$$

There were high correlations between students' ISTEP+ and MAP scores from the same testing intervals, ranging from 0.84 to 0.89 in reading and 0.87 to 0.96 in math (table C1).[13] Further, correlations of students' ISTEP+ scores over time were almost identical to the year-to-year correlations of students' ISTEP+ and MAP scores, which would not be expected if there were large differences in content between the two tests. The standard deviations of value-added estimates by subject, test, and type are in table C2.

Table C1. Student-level correlations of ISTEP+ and MAP scores, 2005/06–2010/11

Testing interval and tests compared[a]	Reading		Math	
	Corr. (r)[b]	*n*	Corr. (r)[b]	*n*
Fall test scores				
$ISTEP_t$ and MAP_t	0.87	6,511	0.92	6,529
$ISTEP_t$ and $ISTEP_{t+1}$	0.89	6,511	0.90	6,529
MAP_t and MAP_{t+1}	0.89	15,009	0.96	15,019
$ISTEP_t$ and MAP_{t+1}	0.84	6,511	0.87	6,529
Spring test scores				
$ISTEP_t$ and MAP_t	0.85	7,212	0.93	7,068
$ISTEP_t$ and $ISTEP_{t+1}$	0.85	7,212	0.89	7,068
MAP_t and MAP_{t+1}	0.87	14,674	0.93	14,645
$ISTEP_t$ and MAP_{t+1}	0.84	7,212	0.89	7,068

ISTEP+ is the Indiana Statewide Testing for Educational Progress Plus. MAP is the Measures of Academic Progress assessment that was administered in the same schools and during the same period as the ISTEP+.

a. The subscript *t* indicates the test interval in a given year, and the subscript *t*+1 indicates the test interval in the subsequent year.

b. Values are Pearson correlation coefficients (r), which are disattenuated of measurement error using the internal consistency reliability coefficients of the ISTEP+ and MAP (see discussion in appendix).

Source: Authors' calculations based on data from the Indiana Department of Education and Northwest Evaluation Association.

Table C2. Standard deviations of estimates of teacher value-added by subject, test type, and testing interval, 2005/06–2010/11

	Fall-to-fall ISTEP+	Spring-to spring ISTEP+	Fall-to-fall MAP	Spring-to spring MAP	Number of teachers	Number of students
Reading						
Year						
2005/06	0.088	na	0.063	[a]	263	4,113
2006/07	0.048	na	0.081	0.094	156	2,398
2007/08	—	na	0.077	0.144	134	2,114
2008/09	na	—	0.127	0.130	190	2,950
2009/10	na	0.176	0.145	0.169	220	3,434
2010/11	na	0.217	[a]	0.176	186	3,778
Grade						
4	0.080	0.196	0.068	0.134	629	10,296
5	0.070	0.195	0.137	0.163	520	8,491
Total	0.075	0.195	0.105	0.148	1,149	18,787
Math						
Year						
2005/06	0.190	na	0.137	[a]	263	4,142
2006/07	0.187	na	0.110	0.226	151	2,387
2007/08	—	na	0.085	0.170	141	2,243
2008/09	na	—	0.141	0.163	190	2,947
2009/10	na	0.247	0.085	0.233	209	3,300
2010/11	na	0.233	[a]	0.211	189	3,768
Grade						
4	0.188	0.252	0.116	0.193	628	10,271
5	0.191	0.226	0.119	0.215	515	8,516
Total	0.189	0.240	0.117	0.204	1,143	18,787

ISTEP+ is the Indiana Statewide Testing for Educational Progress Plus. MAP is the Measures of Academic Progress assessment that was administered in the same schools and during the same period as the ISTEP+.

na is not applicable because tests were administered during a different interval that year.

— is not available because in 2008/09 the administration of the ISTEP+ switched from the fall to the spring, so ISTEP+ scores were not available for adjacent spring-to-spring or fall-to-fall testing intervals.

a. Spring-to-spring MAP estimates for 2005/06 are not available because the research team did not have access to MAP pretest scores for spring 2005. Fall-to-fall MAP estimates for 2010/11 are not available because the research team did not have access to MAP posttest scores for fall of 2011.

Source: Authors' calculations based on data from the Indiana Department of Education and Northwest Evaluation Association.

Notes

1. State laws include Illinois's Public Act 96–0861 ("Performance Evaluation Reform Act," signed into law June 2011), Indiana's Senate Bill 1 (signed into law April 2011), Michigan's Public Act 205 of 2009 (enacted January 2010), Minnesota's revised statute 122A.40 subd. 8 (enacted July 2011), and Ohio's HB 153 (enacted 2009, modified 3319.112 OCR). Wisconsin's Department of Public Instruction (2011) recommends that districts develop systematic teacher evaluation systems in which student growth is a major component.
2. Several states and districts are setting up frameworks and piloting teacher evaluation systems that incorporate results from value-added models. The Minnesota Department of Education educator evaluation model and the Chicago Public Schools Reach Students evaluation model are two notable examples (Minnesota Department of Education, 2012; Chicago Public Schools, 2012). Other examples include Indiana Department of Education (n.d.), Michigan Council for Educator Effectiveness (2012), and Ohio Department of Education (n.d.).
3. For instance, the Ohio Department of Education (2012) requires districts to use multiple assessments to measure student growth. Districts must use state-provided value-added measures based on state test scores in reading and math for grades 4–8. The department approved a list of vendor assessments that districts can choose from to provide growth measures for teachers in grades not covered by the state test or to augment the value-added measures for teachers in grades covered by the state test. MAP was one of the assessments approved by the department. In the state's guidelines for teacher evaluations 50 percent of the final summative ratings for teachers are to be based on student growth measures.
4. The design of the ISTEP+ is similar to that of the tests used in other Midwest Region states, such as Illinois and Ohio. The tests consist of roughly 50 items, and scores are expressed on a vertical scale across grades 3–8. The timing of the ISTEP+ test administration is also similar to that in other states (see, for example, Illinois State Board of Education, 2011b).
5. Sass (2008) found a correlation of 0.48 when comparing value-added estimates from different tests in Florida. Using data on grade 4 and 5 teachers in Houston, Texas, Corcoran et al. (2011) reported correlations of single-year value-added estimates from the Texas Assessment of Academic Skills/Texas Assessment of Knowledge and Skills and the Stanford Achievement Test (SAT 10) ranging from 0.463 to 0.475 in reading and from 0.528 to 0.542 in math.
6. For example, the average value-added estimate in the top quintile on the ISTEP+ spring-to-spring results is 1.15 standard deviations above the average for estimates in the third quintile ((0.218 – 0.006) / –0.195 = 1.149). According to Schochet and Chiang (2010), a difference this large translates to a third of a school year's worth of academic progress. The average value-added estimate in the top quintile was 0.218 (in *z*-score units), compared with –0.006 in the third quintile. The difference between the two estimates ([0.218 – 0.006] / –0.195 = 1.149) was divided by the standard deviation of value-added estimates for spring-to-spring ISTEP+ reading to arrive at 1.15.
7. The attenuated correlations reported by Ho and Kane (2013) ranged from 0.365 to 0.469, while the disattenuated correlations ranged from 0.733 to 0.914. The attenuated correlations were referenced in this report for consistency with the correlations of ISTEP+ and MAP value-added estimates, which were also attenuated.
8. District D's fall MAP window ends three weeks prior to the fall ISTEP+ test window.

9. Because the model is easy to specify and does not require propriety software, districts can replicate this analysis using standard statistical computing packages. And because the model does not require test scores to be linked across grades with a vertical scale, it is well suited for districts that use both norm-referenced and state tests.
10. Unique identifiers for teachers are not consistent across years to allow for longitudinal analysis.
11. Results are similar when school fixed effects are included in the model. The advantage of school fixed effects is that they will control the effects of all unobserved factors that differ systematically across schools, but they do not vary within schools. The drawback is that they will also capture systematic differences in average teacher quality across schools; therefore, estimates of value-added from teachers in different schools are not directly comparable.
12. Lockwood et al. (2007), Papay (2011), and Corcoran et al. (2011) use random effects in their comparisons of estimates of teacher value-added from different tests. The analysis was also done using teacher fixed effects and removing classroom-level aggregates. The fixed-effect model used an errors-in-variables specification to make an upward adjustment to the coefficient on the pretest scores based on their reliability. This is designed to counter the effects of attenuation caused by measurement error in the pretest.
13. Anderson, Alonzo, and Tindal (2010) find correlations of approximately 0.70 between students' math scores on the easyCBM online benchmark assessments and the Oregon state test. Papay (2011) finds correlations of approximately 0.80 across students' scores on three different reading tests.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*(1), 95–135.

Anderson, D., Alonzo, J., & Tindal, G. (2010). *easyCBM® mathematics criterion related validity evidence: Oregon state test*. Eugene, OR: Behavioral Research and Teaching. http://eric.ed.gov/?id=ED531666

Ballou, D. (2005). Value-added assessments: Lessons from Tennessee. In R. Lissitz (Ed.), *Value-added models in education: Theory and application* (pp. 272–297). Maple Grove, MN: JAM Press.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008). *Teacher preparation and student achievement* (NBER Working Paper No. 14314). Cambridge, MA: National Bureau of Economic Research. http://eric.ed.gov/?id=ED502819

Chicago Public Schools. (2011). *Recognizing educators advancing Chicago students*. Retrieved December 18, 2012, from http://www.cps.edu/Pages/reachstudents.aspx

Chicago Public Schools. (2012). *Information on the ISAT value-added metric*. Retrieved November 10, 2012, from http://www.cps.edu/Pages/valueadded.aspx

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Corcoran, S. P., Jennings, J. L., & Beveridge, A. A. (2011). *Teacher effectiveness on high- and low-stakes tests* (Working Paper). New York: New York University.

Cronin, J. (2005). *NWEA reliability and validity estimates: Achievement level tests and Measures of Academic Progress*. Portland, OR: Northwest Evaluation Association.

Goldhaber, D., & Hansen, M. (2010). *Is it just a bad class? Assessing the stability of measured teacher performance* (CEDR Working Paper No. 2010–3). Seattle, WA: Center for Education Data & Research. http://eric.ed.gov/?id=ED537151

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58*(1), 78–80.

Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M., Humez, A., Litke, E., & Lynch, K. (2012). Validating arguments for observational instruments: Attending to multiple sources of variation. *Educational Assessment,* 17(2–3), 88–106. http://eric.ed.gov/?id=EJ980534

Ho, A.D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieval date for NWEA citation: January 15, 2011. http://eric.ed.gov/?id=ED540957

Illinois State Board of Education. (2011a). *2011 teacher/principal evaluation systems—Final collection*. Springfield, IL: Author. Retrieved December 1, 2011, from http://www.isbe.net/peac/pdf/survey/teacher_prin_eval_survey11_final.pdf

Illinois State Board of Education. (2011b). *Student assessment—Illinois Standards Achievement Test*. Retrieved December 1, 2011, from www.isbe.state.il.us./assessment/isat.htm

Indiana Department of Education. (2011). *2010–2011 ISTEP+ program manual*. Retrieved March 1, 2011, from http://www.doe.in.gov/sites/default/files/assessment/istep-program-manual.pdf

Indiana Department of Education. (n.d.). *Educator effectiveness*. Retrieved December 5, 2012, from http://www.doe.in.gov/improvement/educator-effectiveness

Isenberg, E., & Hock, H. (2011). *Design of value-added models for IMPACT and TEAM in DC public schools, 2010–2011 school year. Final report*. Princeton, NJ: Mathematica Policy Research. http://eric.ed.gov/?id=ED521120

Kane, T., & Staiger, D. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (NBER Working Paper No. 14607). Cambridge, MA: National Bureau for Economic Research. http://eric.ed.gov/?id=ED503840

Kingsbury, G. (2003). A long-term study of the stability of item parameter estimates. Paper presented at the 2003 annual meeting of the American Educational Research Association, April 21–25, Chicago.

Koedel, C., & Betts, J. (2007). *Re-examining the role of teacher quality in the educational production function* (Working Paper No. 0708). Columbia, MO: University of Missouri. http://eric.ed.gov/?id=ED510527

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. http://eric.ed.gov/?id=EJ872464

Lefgren, L. & Sims, D. (2012). Using subject test scores efficiently to predict teacher value-added. *Educational Evaluation and Policy Analysis, 34,* 109–121. http://eric.ed.gov/?id=EJ956816

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.-N., & Martinez, F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44,* 1, 47–67. http://eric.ed.gov/?id=EJ763893

McCaffrey, D. F., Lockwood, J. R., Koretz, D., & Hamilton, L. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation. http://eric.ed.gov/?id=ED529961

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67. http://eric.ed.gov/?id=EJ727500

McCaffrey, D. F. Sass, T., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572–606. http://eric.ed.gov/?id=EJ863346

Michigan Council for Educator Effectiveness. (2012). *Interim progress report*. Lansing, MI: Author. Retrieved May 15, 2012, from http://www.michigan.gov/documents/mde/Interim_Progress_Report_MCEE_383698_7.PDF

Milanowski, A. (2011). *Resolving some issues in using value-added measures of productivity for school and teacher incentives: Ideas from technical assistance and TIF grantee experience*. Madison, WI: Center for Educator Compensation Reform. http://eric.ed.gov/?id=ED533044

Milanowski, A., Heneman, H., & Kimball, S. (2011). *Teaching assessment for teacher human capital management: Learning from the current state of the art* (WCER Working Paper No. 2011–12). http://eric.ed.gov/?id=ED517293

Minnesota Department of Education. (2012). *Educator evaluation*. Retrieved May 15, 2012, from http://education.state.mn.us/MDE/EdExc/EducEval/index.html

Northwest Evaluation Association. (2011a). Growth Research Database. Retrieved January 15, 2011, from http://www.kingsburycenter.org/our-data/grd-data

Northwest Evaluation Association. (2011b). *The RIT scale*. Portland, OR: Author. Retrieved July 10, 2011, from http://www.nwea.org/support/article/532/rit-scale

Ohio Department of Education. (2012). *Student growth measures for teacher evaluation*. Retrieved May 15, 2012, from http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDetail.aspx?page=3&TopicRelationID=1230&ContentID=125742

Ohio Department of Education. (n.d.). *Educators*. Retrieved May 15, 2012, from http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDetail.aspx?page=3&TopicRelationID=1230&ContentID=125739&Content=130822

Papay, J. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *The American Educational Research Journal, 48*(1), 163–193. http://eric.ed.gov/?id=EJ911163

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6(1), 15–32.

Sass, T. (2008). *The stability of value-added measures of teacher quality and implications for teacher compensation policy*. Washington, DC: The Urban Institute.

Sass, T., Hannaway, J., Xu, Z., Figlio, D., & Feng, L. (2012). Value added of teachers in high-poverty schools and lower-poverty schools. *Journal of Urban Economics, 72*(2), 104–122.

Schug, M., & Niederjohn, M. (2009). *Value added testing: Improving state testing and teacher compensation in Wisconsin*. Hartland, WI: Wisconsin Policy Research Institute. Retrieved May 21, 2012, from http://www.wpri.org/Reports/Volume22/Vol22No4/Vol22No4.html

Value-Added Research Center. (2012a). *Developing a value-added system with Madison Metropolitan School District*. Retrieved May 21, 2012, from http://varc.wceruw.org/Projects/developing_va_mmsd.php

Value-Added Research Center. (2012b). *Milwaukee classroom value-added initiative*. Retrieved May 21, 2012, from http://www.wcer.wisc.edu/projects/projects.php?project_num=476

Wisconsin Department of Public Instruction. (2011). *Wisconsin framework for educator effectiveness: Design team report & recommendations*. Retrieved May 21, 2012, from http://ee.dpi.wi.gov/files/ee/pdf/EEFinalDesignTeamReport-11%207%2011.pdf