NATIONAL
CENTER for ANALYSIS of LONGITUDINAL DATA in EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

The Common Core Conundrum: To What Extent Should We Worry That Changes to Assessments and Standards Will Affect Test-Based Measures of Teacher Performance?

Ben Backes
James Cowan
Dan Goldhaber
Cory Koedel
Luke Miller
Zeyu Xu

WORKING PAPER 152 • February 2016

The Common Core Conundrum: To What Extent Should We Worry That Changes to Assessments and Standards Will Affect Test-Based Measures of Teacher Performance?

Ben Backes
American Institutes for Research

James Cowan
American Institutes for Research

Dan Goldhaber
American Institutes for Research and University of Washington

Cory Koedel
University of Missouri

Luke Miller
University of Virginia

Zeyu Xu
American Institutes for Research

Contents

Acknowledgements

This research was made possible in part by generous support from the National Center for the Analysis of Longitudinal Data in Education Research (CALDER). Research supported by CALDER is funded through Grant R305C120008 to the American Institutes for Research from the Institute of Education Sciences, U.S. Department of Education.

We are grateful to feedback from Sean Corcoran, Lars Lefgren, and discussants at APPAM. We also thank our state partners for providing data access. Melanie Rucinski provided excellent research assistance.

CALDER • American Institutes for Research
1000 Thomas Jefferson Street N.W., Washington, D.C. 20007
202-403-5796 • www.caldercenter.org

Suggested Citation: Backes, B., Cowan, J., Goldhaver, D., Koedel, C., Miller, L., Xu, Z. The Common Core Conundrum: To What Extent Should We Worry That Changes to Assessments and Standards Will Affect Test-Based Measures of Teacher Performance? Working Paper 152. *National Center for Analysis of Longitudinal Data in Education Researc*h.

The Common Core Conundrum: To What Extent Should We Worry That Changes to Assessments and Standards Will Affect Test-Based Measures of Teacher Performance?

Ben Backes, James Cowan, Dan Goldhaber, Cory Koedel, Luke Miller, Zeyu Xu

CALDER Working Paper No. 152
February 2016

Abstract

Using administrative longitudinal data from five states, we study how value-added measures of teacher performance are affected by changes in state standards and assessments. We first document the stability of teachers' value-added rankings during transitions to new standard and assessment regimes and compare our findings to stability during stable standard and assessment regimes. We also examine the predictive validity of value-added estimates during nontransition years over transition-year student achievement. In most cases we find that measures of teacher value added are similarly stable in transition years and nontransition years. Moreover, there is no evidence that the level of disadvantage of students taught disproportionately influences teacher rankings in transition years relative to stable years. In the states we study, student achievement in math can consistently be forecasted accurately—although not perfectly—using value-added estimates for teachers during stable standards and assessment regimes. There was somewhat less consistency in reading, because we find cases where test transitions significantly reduced forecasting accuracy.

I. Introduction

Changes to state educational standards and assessment regimes are common. For instance, in the five states studied in this paper, there have been 14 standards or assessment changes in math and reading since 2000. Indeed, in some states, the revision of standards and assessments is routine.[1] The implications of such changes for teacher evaluation are now receiving increasing scrutiny due to the increased use of student test-based measures of teacher effectiveness (value added) and the widespread implementation of the Common Core State Standards (CCSS).[2] The confluence of test-based teacher evaluations and the transition to CCSS has generated considerable controversy among educators and policy makers.[3]

A central objection to the implementation of these two initiatives is the idea that it is unfair to hold teachers accountable for results on the initial year of a new assessment that is designed to be a more rigorous test of student learning.[4] Some policy makers and practitioners, and most prominently teachers unions, have argued that teachers need more time to develop lessons and learn about the new tests before being held accountable for their students' performance on them.[5] In response to these concerns, then-Secretary of Education Arne Duncan granted a one-year moratorium on the use of test-

[1] North Carolina, one of the sites for this study, revised its standards and associated assessments on a recurring five-year schedule, with a previous revision described as a "drastic change in the curriculum" (Bazemore et al., 2006). The state typically did not use the initial year of an assessment to count for student grades (Helms, 2013).

[2] Two important pillars in Race to the Top, a multibillion-dollar effort by the federal government to encourage reform and improvement in America's public schools, are performance-based reviews for teachers and the implementation of CCSS.

[3] "Policymakers and educators alike are grappling with the reality that the inputs (such as state tests) used in accountability measures are changing—and they are often resistant to using student test data to trigger negative consequences usually associated with poor performance. Of particular concern is how to calculate growth as students transition from one exam to another and what to do about growth-based accountability and evaluation systems in the interim." Sears, Victoria. "State Accountability in the Transition to Common Core." May 2014. *The Fordham Institute.*

[4] For example, see Chang, Kenneth. "With Common Core, Fewer Topics but Covered More Rigorously." 3 September 2013, *The New York Times*, D2.

[5] For instance, AFT president Randi Weingarten argued that "the tests are evaluating skills and content these students haven't yet been taught." Source: Rose, Mike. "AFT calls for moratorium on Common Core consequences." *AFT News*, April 2013.

based teacher evaluations in 2014.[6] But are these concerns well-founded? While it is not possible to know *a priori* the extent to which any specific standards or testing change will result in meaningful impacts on judgments about teacher performance, the fact that standards and assessment changes are not new affords the opportunity to assess how past changes have affected value-added measures of teacher effectiveness.[7]

A handful of recent studies have investigated the stability of value-added measures over time, across calculation methods, across schools, and across testing instruments within the same curricular regime.[8] These studies typically find that a sizeable portion of a teacher's performance persists over time and in different schools, but not always across testing instruments. Less is known, however, about whether successful teachers under one set of standards and one testing regime continue to be successful after the implementation of a new regime. We address this issue in this paper, reporting on research assessing the extent to which student test-based measures of teacher performance are affected by standards and testing changes. Specifically, we use longitudinal data from Kentucky, Massachusetts, New York City, North Carolina, and Washington, each of which previously revised its standards and student assessments, to explore the reliability and stability of teacher value added during transitions across standards and testing regimes (including, in some cases, transitions to CCSS).

[6] Announcement: http://www.ed.gov/blog/2014/08/a-back-to-school-conversation-with-teachers-and-school-leaders/

[7] In principle, standards and curricula are distinct in that standards are expectations about what a student will learn in a given school year and curricula are means by which that learning is accomplished. However, disentangling the two in order to estimate the effect of changing one or the other would not be simple. For example, in North Carolina, the *Standard Course of Study* refers to both the state's standards and curriculum ("The 1998 Mathematics Standard Course of Study and North Carolina Mathematics Tests." *Public Schools of North Carolina*, October 2000); furthermore, adopting CCSS entails shifting to a new curriculum aligned with the new standards. As a result, we focus on test changes and standards changes, but for practical purposes we treat standards changes and curriculum changes as equivalent.

[8] For stability over time, see McCaffrey et al. (2009) and Goldhaber and Hansen (2013). For stability across calculation methods, see Ehlert et al. (2014) and Goldhaber et al. (2013). For stability across schools, see Xu et al. (2012), Glazerman et al. (2013), and Chetty et al. (2014). And for stability across testing instruments, see Lockwood et al. (2007), Corcoran et al. (2011), and Papay (2011).

The standards and testing transitions in the sites we study occurred within the context of a wide range of assessment and evaluation policies.[9] We study two states that began assessing the CCSS before the introduction of the tests offered by the Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced Assessment (SBAC) consortia. In one of our states, Massachusetts, districts in one year had a choice of whether to adopt PARCC or continue to use their existing CCSS-aligned test. We study three states that adopted new or revised learning standards before adopting the CCSS and two states that revised their assessments without altering the underlying learning standards. The variation in these policy changes reflects the diverse experiences of states transitioning to the CCSS standards and assessments. As of June 2015, 43 states and the District of Columbia had at least partially adopted the CCSS (Ujifusa, 2015). However, only 20 states and the District of Columbia currently belong to one of the two testing consortia (Zernike, 2015).

We assess two related concerns about the use of value-added models during testing transition years. First, we assess the extent to which teachers' value-added rankings change as the result of a new standards or assessment regime relative to rankings changes holding the regime fixed by estimating and comparing the stability of teacher rankings over time when there is and is not a change in the standards and assessments. Second, we draw on the approach of Chetty, Friedman, and Rockoff (2014) and Bacher-Hicks, Kane, and Staiger (2014) to assess the degree to which value added in stable years predicts student achievement during transitions.

We find little evidence that measures of teacher effectiveness are any less stable in transition years than in nontransition years in three of the five states for reading, and for all five states in math. When measuring the share of teachers in the top or bottom deciles of the distribution of value added

[9] The standards and assessment changes we observe occur in a variety of accountability settings due to the length of our panel in some states. It is possible that a state's accountability regime affects how teachers react to standards and assessment changes, but in general the transitions we observe do not coincide with the adoption of new accountability policies.

who remain in the same decile the following year, the likelihood of decile persistence in transition years in math is similar to stable years, with one exception (middle school in Massachusetts, with a fall of 10 percentage points for the top decile and 7 for the bottom). For reading, at two sites—Kentucky and Massachusetts—we observe drops in classification consistency in transition years, with the maximum decline on the order of 13 percentage points in middle school in Massachusetts. We also find that in some cases, volatility actually decreases with the transition to a new regime. There is also no evidence that volatility in value added in transition years is associated with the degree of disadvantage of the students to which teachers are assigned.

Student achievement during transition years can also generally be forecasted accurately, although not perfectly, by teacher value added from stable standards and assessment regimes. In most cases, we estimate deviations from our forecasts during transition years ranging from 0 to 10 percent, with upper bound estimates as high as 20 percent. Two notable exceptions occur in our analysis of reading value added, where in Kentucky and Massachusetts we find significantly larger forecast deviations–roughly 40 percent.

A general takeaway from our study is that value-added measures of teacher quality provide useful information about educator efficacy, even during standard and assessment transitions. Although our imperfect forecasts during transitions are consistent with some loss of information, much of the informational content of value added clearly persists, especially in math. Our findings are also consistent with other recent research (Lefgren & Sims, 2012; Goldhaber & Hansen, 2013) in that we find reading value added is less informationally robust than math value added.

In the two instances where reading value added carries through the transition particularly poorly, we are unable to identify clear mechanisms driving the instability we find. Thus, while our findings point broadly toward value added continuing to be a useful measure of teacher effectiveness during transitions, and particularly so in math, our investigation suggests that *ex ante* it may be difficult

to identify situations where the informational content of value added will change meaningfully during a transition. This uncertainty may contribute to calls to halt test-based assessments of teachers during transitions, especially assessments with high stakes attached. This might be appropriate if other means of assessing teacher performance prove to be more informationally robust across transitions, or if the cost of this informational uncertainty to teachers (psychological or otherwise) is deemed too high. However, we note that such a policy choice would be equivalent to treating the informational content of value-added measures during transition years as null, which from a purely analytic perspective is not supported by our analysis. Even in the worst transitional cases we identify, our results indicate that teacher evaluations are improved by value-added measures, assuming that a part of a teacher's performance ought to be his or her ability to improve students' achievement.

II. Background

A. Evidence on the Reliability and Stability of Measured Teacher Performance Over Time

Estimates of teacher value added typically assume that there is a persistent component of value added that is constant for a teacher over time (e.g., Kane & Staiger, 2008; McCaffrey et al., 2009; Goldhaber & Hansen, 2013; Chetty et al., 2014 and distinct from nonpersistent changes in teacher performance and from sampling errors.[10] There is evidence that the persistent component of teacher effectiveness is stable across settings. For instance, Chetty et al. (2014) and Bacher-Hicks et al. (2014) find that VA estimates for teachers who switch schools and grades exhibit no forecast bias in two different large school districts. Xu et al. (2012) find little evidence of a change in teachers' measured effectiveness before and after switching schools in either North Carolina or Florida. Similarly, findings from the Talent Transfer Initiative (Glazerman et al., 2013; Glazerman & Protik, 2015) indicate that

[10] Studies find that the stability of value added increases significantly when additional years of data are used to estimate it (McCaffrey et al., 2009; Goldhaber & Hansen, 2013).

teachers continue to have positive effects on student achievement in math and reading when selected to transfer to high-poverty schools.

Recent work has also documented nonpersistent changes in teacher effectiveness distinct from measurement error. Goldhaber and Hansen (2013) find that, although teacher performance is not completely constant over time, its stability is in line with that of performance measures from other occupations. McCaffrey et al. (2009) find that time-varying teacher effects constitute about half of the variation not due to noise in elementary school teacher value added. Chetty et al. (2014) account for the "drift" in estimated teacher value added caused by real changes in performance over time by imposing some light structure in their analysis—in particular, they model a stationary process where mean teacher quality does not vary over time and the covariance between value added estimates in two different years depends only on the amount of time between those years.

At least two factors govern how individual teachers will fare after a curriculum and assessment regime change. The first is adaptability. For example, it could be the case that effective teachers are more capable in general and thus better able to adapt to new standards. Alternatively, it may be the case that effective teachers are identified as such because they have built up specialized knowledge under a given curriculum and assessment regime, and thus will be harmed by a switch. Second, research shows that teachers are differentially effective at teaching various components of a given subject (Lockwood et al., 2007; Papay, 2011) and have different measured performance across tests that may differ in focus (Corcoran et al., 2011; Papay, 2011). In particular, Papay (2011) argues that although correlations between value-added estimates derived from different tests are moderately high, with a range of 0.15 to 0.58, the resulting differences are sufficient to induce considerable instability to teacher rankings. Thus, curriculum and assessment changes that emphasize particular tasks could favor some teachers over others and lead to ranking changes among the workforce. To our knowledge, however,

there is no direct evidence on how these and other potential factors act together to influence measured teacher performance during the transition to a new curriculum and assessment regime.

B. Standards and Testing Changes Across States

The implementation of the CCSS has engendered a great deal of discussion about the implications of the new standards, curriculum, and tests but, as noted above, it is actually quite common for states to revise their standards and assessments. Previous regime changes can be used to assess how teacher rankings are affected during standard and assessment regime changes, which can provide insight regarding states' CCSS changes. Below we briefly describe the standard and assessments changes in the states that are the focus of our study (see Appendix A for more detail on these changes).

Kentucky: In 2009, Kentucky began to develop a new set of standards and accompanying assessments in all subjects to align with the CCSS. Kentucky adopted the new CCSS-aligned standards in 2010, becoming the first state to do so. During the 2010-2011 school year, district leadership teams constructed student learning targets from the standards, which were then shared with teachers.[11] The new standards were taught for the first time in the 2011-2012 school year, and students took the new assessments for the first time in the spring of 2012.

The assessment in Kentucky before the adoption of the new test in 2012 has been shown to have a maximum score that is attained by many students (Innes, 2009; Koretz et al., 2014), which we illustrate in Figure A1. As shown in Section IV, estimates of teacher effectiveness in Kentucky are among the most volatile in the transition year, and the properties of this prior exam could be a contributing factor. The skewness for the old Kentucky test, however, is about -0.35, which, based on results from Koedel and Betts (2010), should not be enough to cause substantial bias.[12]

[11] Kentucky Department of Education. The Facts about Kentucky's Core Academic Standards in English Language Arts and Mathematics. http://education.ky.gov/comm/UL/Documents/Facts%20about%20the%20KCAS%202014.pdf

[12] As described in the appendix, we implement an alternative specification by probit transforming all pretest and posttest measures in Kentucky. However, this transformation makes very little difference in the results, with the

Massachusetts: Massachusetts formally adopted learning standards that incorporated the CCSS in math and reading in 2010. The prior standards were adopted in 2000 and revised in 2004. The complete set of grade-level standards was first assessed in 2006 (Massachusetts Department of Elementary and Secondary Education, 2004a,b). Following the adoption of the CCSS, Massachusetts used an assessment focusing on areas common to the two sets of standards in 2012, and began assessment of the new math standards in 2013.[13, 14] In English and Language Arts,[15] Massachusetts moved to the new standards during the 2013 testing cycle.[16]

In 2015, the state offered each district the choice of whether to continue administering the Massachusetts Comprehensive Assessment System (MCAS) or to switch to the Partnership for Assessment of Readiness for College and Careers (PARCC), with about half of districts choosing PARCC. Thus, the 2015 MCAS districts did not have an assessment or standards change, whereas the 2015 PARCC districts had a new assessment under the same standards.[17] In addition, about half of the PARCC districts administered an online test in place of a paper test. Results are generally similar when restricting the sample of PARCC districts to the paper-only districts (see discussion in Appendix A).

---

exception of the estimate of forecast bias in the reading assessment (Table 6, column 5), which shrinks from 42 percent to 28 percent when performing the transformation. In addition, the skewness of the test in Kentucky is not substantially larger than what is observed in some of the other sites in this paper.

[13] Massachusetts Department of Elementary and Secondary Education. (2014, March 18). Assessment Transition Plans - Massachusetts Comprehensive Assessment System. Retrieved November 24, 2015, from http://www.doe.mass.edu/mcas/transition/?section=math3-8. Massachusetts Department of Elementary and Secondary Education. (2014, March 18).

[14] Although Massachusetts adopted the CCSS, the state did not immediately adopt one of the two main standardized tests associated with the standards. Massachusetts participated in a field test of PARCC in 2014 and allowed school districts to choose between the PARCC and the existing state test in 2015. In November 2015, the State Board of Education voted to modify the existing state test rather than switch to the PARCC assessment (Fox, 2015; Zernike, 2015).

[15] For convenience, we use "English and Language Arts" and "Reading" interchangeably throughout the paper.

[16] Assessment Transition Plans: Assessment Transition from 2001/2004 MA English Language Arts (ELA) Framework to 2011 MA Curriculum Framework for English Language Arts & Literacy - Massachusetts Comprehensive Assessment System. Retrieved November 24, 2015, from http://www.doe.mass.edu/mcas/transition/?section=ela

[17] When comparing teachers in districts that would eventually choose PARCC to those that would remain with MCAS, we do not observe any differences in estimated value added in years before 2015.

New York City: Before 2006, New York City (NYC) used its own test for Grades 3-8. In 2006, statewide testing was introduced in response to No Child Left Behind; these tests were accompanied by a standards change and replaced the NYC district test. The state test was field tested in 2005 and first administered early in the 2006 calendar year.[18] To assist districts in adopting the state standards, the state shared a toolkit to support districts in aligning their curriculum to the new standards.[19]

North Carolina: Before the adoption of the CCSS, curriculum revisions in North Carolina operated on a five-year schedule and were planned well in advance. For example, for the transition to mathematics Edition 2, which was ultimately first implemented during the 1999-2000 school year, a new K-12 curriculum was adopted in May 1998. The new curriculum was used for instructional planning and textbook selection during the year before implementation. Field test items corresponding to the new curriculum were included in end-of-grade tests in the spring of 2000 and the new tests were introduced formally in 2001.[20] All North Carolina standards and assessment changes that we study in North Carolina occurred simultaneously—that is, new assessments are always accompanied by new standards, with teachers being exposed to the new standards well ahead of the new assessments.[21]

Washington: We examine assessment and standards changes in Washington since the state began annual statewide standardized testing in Grades 3-8 in the spring of 2006. These new state assessments reflected a set of learning standards for each of grades K-10 that were introduced in 2004 (Office of the Superintendent of Public Instruction, 2004a, 2004b).[22] In July 2008, Washington released

[18] Introduction to the Grades 3-8 Testing Program in English Language Arts and Mathematics. http://www.scotiaglenvilleschools.org/parentinformation/38intro.pdf

[19] Kline, Michelle. New York State Education Department Forum on NYS Learning Standard for Mathematics. Link (not displayed due to length).

[20] "The 1998 Mathematics Standard Course of Study and North Carolina Mathematics Tests." *Public Schools of North Carolina*, October 2000.

[21] The transition to CCSS in North Carolina took place on a similar timetable but is not used in this study. "Public Schools of North Carolina Development Timeline." Accessed at http://www.ncpublicschools.org/core-explained/timeline/

[22] Prior to 2006, Washington relied on grade span testing in grades 4, 7, and 10, and the previous state learning standards specified benchmarks for those grades only.

new math standards (Office of the Superintendent of Public Instruction, 2008), which were formally assessed for the first time during the 2009-2010 school year. At the same time, the format of the state assessment changed in both math and reading, although only the math assessment change was accompanied by a change in state standards.[23]

Table 1 displays a list of assessment changes under consideration for this study.[24] The reading assessment changes in New York City and Washington and the 2015 PARCC transition for some districts in Massachusetts are the only cases where an assessment change was not accompanied by a standards change. For states where both elementary and middle grades are available, we analyze results separately to allow for different patterns across school types.

III. Data and Analytic Approach

A. Data

We use administrative data covering different time periods in Kentucky, Massachusetts, New York City, North Carolina, and Washington. Although each site's data are unique in the sense that they span different years and grades, and they contain slightly different information about student background characteristics, they also share commonalities in terms of their general content and structure. For instance, each state provides the information needed to link students to their teachers and track teachers over time.[25] In addition, we observe demographic information such as race, gender, and free or reduced-price lunch (FRL) eligibility.

[23] Although we do not consider Washington's experience with the CCSS here, the state began partial implementation in classrooms during the 2012-2013 school year and assessment during the 2014-2015 school year.
[24] Detailed information about test and curriculum changes in each state is contained in the Appendix.
[25] Nonrandom attrition of teachers in response to assessment or standards changes could potentially affect results. In results available from the authors, we measure for each year the share of teachers in the dataset who are again observed in the following year. We do not find any evidence that teachers were more likely to exit in transition years.

Table 2 displays information for each state about the years and grades available, demographic information included, and number of students and teachers. Detailed information about the construction of analytic samples in each state can be found in Appendix A.

B. Estimating Teacher Value Added

We estimate a standard value added model where current achievement can be expressed as:

$$A_{ijt} = \lambda A_{it-1} + \alpha X_{it} + \beta_{jt} T_{jt} + \eta_{ijt} \qquad (1)$$

where A_{ijt} denotes achievement of student *i* taught by teacher *j* in year t, A_{it-1} prior achievement, X_{it} demographic controls (which vary by state), and T_{jt} a teacher fixed effect. The coefficients on these fixed effects are taken as our measures of teacher value added. We control for prior test scores using a cubic polynomial in prior-year scores in math and English as in Chetty et al. (2014). Chetty et al. (2014) provide evidence that controlling for prior-year test scores in this way removes nearly all of the bias from measures of teacher performance. Demographic controls vary among states (see Table 2 for a list of demographic information in each state dataset). We do not control for classroom characteristics; previous research suggests doing so will not affect our findings substantively.[26]

We incorporate prior achievement as a control, rather than using a gainscore model, as this model has been demonstrated to predict future performance well in experimental work (Kane & Staiger, 2008; Kane et al., 2013). Moreover, nonexperimental tests suggest that this model estimates teacher effects with little bias (Chetty et al., 2014; Bacher-Hicks et al., 2014), and simulations find that it is more robust to nonrandom classroom assignment based on previous achievement (Guarino et al., 2015b). As in other studies, we standardize test scores to have a mean of zero and standard deviation of one within

[26] Kane et al. (2013) find that a similar model is predictive of future teacher performance when students are randomly assigned to teachers. Goldhaber et al. (2013) show that estimates of teacher value-added from models that do not include classroom characteristics are highly correlated with estimates from models that do (r=0.99). When we add class average controls to our test for forecast bias, results are very similar.

grade/subject/year. Some investigations into properties of value added use shrinkage estimators, where value added estimates are shrunken towards mean value added to account for sampling error. We do not shrink the estimates of value added obtained from equation (1), noting that Guarino et al. (2015a) show that shrinking estimates does not meaningfully change their reliability and that shrinkage make little difference in the rank ordering of teachers. Consistent with the Guarino et al. findings, the results from our comparative analysis across stable and transitional testing years is not affected substantively by whether or not we explicitly account for estimation error in our estimates of value-added, as shown in Appendix B.[27]

There are at least two reasons we might think teachers' value-added would be sensitive to changes in educational standards and/or testing regimes. To fix ideas, consider a model of the components of estimated teacher effectiveness (used in, e.g., Goldhaber & Hansen, 2013; Winters & Cowen, 2013; and Koedel & Li, 2016, with an additional term for assessment changes):

$$\widehat{VA}_{jt} = q_j + \delta_{jt} + \tau_j^k + \varepsilon_{jt}. \qquad (2)$$

In the above equation, q_j represents a persistent component of teacher quality, δ_{jt} year to year shocks to performance unrelated to testing or standards (for example, classroom match effects), τ_j^k test-specific knowledge of teacher *j* for regime *k*, and ε_{jt} a random error term. We take the first two terms to be invariant to standards and assessment changes, leaving the latter two terms as channels through which estimated value added can be affected by regime changes.

First, per the discussion in Section I, teachers may differ in their ability to adapt to new standards, and their skills may vary across tested subjects and content, which would be reflected by differences in τ_j^k for the same teacher, *j*, under different regimes, *k*. For example, if some teachers excel at teaching

[27] When forecasting value added in section D, the estimates are implicitly shrunken per the procedure that we follow as developed by Chetty, Friedman and Rockoff (2014), and thus no additional *ex post* adjustments for estimation error are made.

more advanced content in particular, they will benefit from switching to a more rigorous set of standards and assessments because the assessment is more sensitive to changes in student learning for students who tend to score at the upper end of the performance distribution. Second, standardized tests typically do not measure student learning at all performance levels with equal accuracy, and changes in the test may affect measured teacher performance in ways that will disproportionately affect teachers depending on the types of students they teach (Koedel, Leatherman, & Parsons, 2012; Stacy, Guarino, Reckase, & Woolridge, 2013; Lockwood & McCaffrey, 2014). An example would be a teacher with a lower-achieving classroom of students who experiences a test change to a more-rigorous assessment. This would lead to less precise measures of achievement for her students, and correspondingly, a less-reliable estimate of her value-added, which would be reflected by an increase in the variance of the error term in equation (2). Relatedly, and more generally, a test change may weaken the fit of the model of student achievement (equation 1), which would create noisier estimates of teacher value added for all teachers. In principle, the results we present below are best viewed as encompassing changes in value added that arise from both of these sources, although in Section V we argue that they are driven primarily by changes to test content rather than issues related to test measurement error.

C. Measuring the Stability of Teacher Performance in Regime Changes

We analyze the stability of teacher rankings by first presenting a descriptive overview and then by performing regression analysis. The descriptive overview displays year-to-year correlations in estimated value added and year-to-year transition likelihoods for teachers ranked in the top and bottom deciles of the value-added distribution. Next we explicitly test for differences in stability by regressing the absolute value of the change in a teacher's percentile rank in the value-added distribution on classroom characteristics and an indicator of whether value added is measured in a transition year.

To define a regime change year, we first denote test scores for individual *i* in time *t* under assessment regime A as A_{it}^{A} and under regime B as A_{it}^{B}. Let *AA* denote value added calculated using pre- and posttests from regime A:

$$AA \equiv VA^{AA} = f\left(A_{it}^{A}, A_{it-1}^{A}\right); \qquad (3)$$

and let *BA* denote value added calculated using a pretest from *A* and posttest from *B* in the initial year after a regime change:

$$BA \equiv VA^{BA} = f\left(A_{it}^{B}, A_{it-1}^{A}\right). \qquad (4)$$

When measuring the correlation in value added for a given teacher in two consecutive years, we define *t* to be a *stable year* when value added in time *t* and time *t-1* are each calculated entirely using tests from the same regime. For example, consider measuring the change in the value-added percentile ranking for a given teacher from time *t-1* to *t*. The following figure illustrates the regimes from which the pre- and posttests used in value-added calculations for North Carolina are taken, with 2001 (spring) being the initial year of a new math assessment. Solid dots in the figure indicate tests from regime *A* and empty dots indicate tests from regime *B*:

	1998	1999	2000	2001	2002	2003	2004
Post-test	●	●	●	○	○	○	○
Pre-test	●	●	●	●	○	○	○

Consider the correlation between estimated teacher value added in 2001 and 2002. The estimates from 2002 are calculated using pre- and posttests from the same regime (regime *B*), but the estimates from 2001 use pretests from the old regime and posttests from the new regime. Estimation in 2001 is during a transition year, and thus we describe the correlation between estimated teacher value added in 2002 and 2001 as a *transitional measurement*. Relating this definition to our current analysis, the key question is whether the year-to-year stability properties of value-added are similar or different when the measurement period is transitional versus stable.

To formalize and extend the descriptive analysis above, we also regress the absolute value of the change in a teacher's percentile ranking between years *t-1* and *t* on classroom characteristics and whether value added was measured during a transition year:

$$|Rank_{jt} - Rank_{jt-1}| = \alpha_0 + \alpha_1 Transition_{t,t-1} + \alpha_2 Char_{jt} + + \varepsilon_{jt}, \quad (5)$$

where $Char_{jt}$ contains teacher *j*'s average class values of %Black, %FRL, and prior test scores, and $Transition_{t,t-1}$ is an indicator equal to one if value added in either *t* or *t-1* was calculated using a transition year (per above, defined as a year with the pretest and posttest taken in different assessment regimes).[28] If α_1 is positive, it would indicate that transition years are associated with increased volatility of teacher rankings relative to the volatility observed during stable curriculum and assessment regimes.

Equation (5) is useful to assess the overall change in the volatility of teacher rankings associated with a regime change, but it is not suited to examining teacher subgroups. It may be the case that volatility increases more for some teachers than others. To examine this possibility, we expand the model in equation (5) as follows:

$$|Rank_{jt} - Rank_{jt-1}| = \beta_0 + \beta_1 Transition_{t,t-1} + \beta_2 Char_{jt} * Transition_{t,t-1} + \beta_3 Char_{jt} + \varepsilon_{jt}. \quad (6)$$

The coefficient vector of interest in equation (6), β_2, measures the extent to which changes in teachers' rankings after a regime shift systematically affect some groups of teachers more than others, as identified by the characteristics of the students they teach. For any particular characteristic, a positive value for β_2 indicates that rankings are more volatile for teachers who teach more students with that characteristic.

D. The Informational Content of Value-Added During Transition Years

[28] An alternative approach in this regression would be to define a transition year to be solely the initial year after a regime change. Results are similar to what we report below if we restrict our attention to the initial year after a regime change.

The preceding analyses can be used to identify the effect of regime changes on the stability of teacher value added. As noted above, this instability can derive from a multitude of factors, and in this subsection, we develop a formal test for changes to the informational content of value added during regime changes. Outside of transition years, it has been well-established that value-added estimates contain useful information about teacher performance (e.g., Bacher-Hicks et al., 2014; Chetty et al., 2014; Kane et al., 2013). Less is known, however, about the validity of value-added estimates during regime changes, and this is an explicit concern raised by critics of measuring transition-year value added, especially for use in teacher evaluation systems. Do transition years change the information value of value added?

To answer this question, we adapt a portion of the analysis performed in Chetty et al. (2014). Specifically, Chetty et al. (2014) provide evidence consistent with teacher value added following a stationary process, and we construct a test to determine whether stationarity in value added is maintained during a standard and assessment transition. The assumption that value added follows a stationary process requires that (1) average teacher quality does not vary over time and (2) the correlation of teacher quality, class shocks, and student shocks across any pair of years depends only on the amount of time that elapses between those years.[29] If the stationarity of value added is upheld through transition years, it suggests that these years are not associated with a fundamental change in the informational content of value added. Alternatively, if stationarity is not maintained through transition years, it suggests that value-added measures from transition years contain different information about performance than value-added measures during nontransition years.

We start with a parallel investigation of stationarity of teacher value-added focusing only on data from stable, nontransition regimes. Specifically, in each year of a stable regime, we forecast teacher value added during year *t,* denoted $\hat{\mu}_{jt}$, using data from all other years within stable standards

[29] See Chetty et al. (2014) for additional details.

and assessment regimes. As described in Chetty et al. (2014), we construct forecasted value added by first estimating the best linear predictor of a teacher's test score residuals in time *t*, using residuals from all other years to establish the time path of residuals, and then predicting each teacher's value added in *t* given the estimated relationship and that teacher's residualized scores in all other years.

We then regress students' residualized test scores (residualized based on pretest scores and demographic information), $\widehat{A_{it}}$, on the forecasted value added of their teachers within stable regimes using the following regression:

$$\widehat{A_{it}} = \alpha_t + \lambda\hat{\mu}_{jt} + \xi_{it}. \qquad (7)$$

Under the stationarity assumption, the OLS regression in equation (7) should yield an estimate of λ that is indistinguishable from one because $\hat{\mu}_{jt}$ is the best linear predictor of A_{it}.

As in Chetty et al. (2014), below we show that our estimates of λ from equation (7) during stable standards and assessment regimes are generally close to one. Having replicated the Chetty et al. (2014) result during stable standards and assessments regimes in the various site under study, we next extend the approach to measure the extent to which value added in transition years can be forecasted with information from stable years.

Note that stationarity implies that teacher value added in year *k*, where $k \neq t$, will predict value added in year *t* the same for all *k* and *t* of fixed distance in time. So, for example, the predictive validity of teacher value added in year 2007-2008 over value added in 2009-2010 will be the same as the predictive validity of value added in year 2011-2012 over value added in 2013-2014. To illustrate how we test the hypothesis of stationarity of value added through transition years, consider a hypothetical standards and assessment regime change over six years. The first three years use old standards and assessments, and the last four years use new standards and assessments. This set-up facilitates three years of stable-regime value-added estimates in the pre period, one transition-year value-added estimate, and three stable-regime value-added estimates in the post period.

	1998	1999	2000	2001	2002	2003	2004
Post-test	●	●	●	○	○	○	○
Pre-test	●	●	●	●	○	○	○
Value-added estimate	S	S	S	T	S	S	S

We test for violations of stationarity during the transition using the following procedure. First, returning to equation (7), we forecast teacher value added during each stable-regime year using data from all stable years and obtain $\hat{\mu}_{jt}$. We store the forecasting coefficients over year *t* value-added for all values of $|t\text{-}k|$, where $k \neq t$. In our example here, this would yield predictive coefficients for $|t\text{-}k|=1$, $|t\text{-}k|=2$, and $|t\text{-}k|=3$. Denote these coefficients for the cross-year correlations estimated from stable regimes as π_1, π_2, and π_3. With these coefficients in hand, and continuing with our example, we can construct a fitted-value measure of predicted teacher value added during the transition year, τ, as follows (with an additional adjustment for class size as in Chetty et al., 2014):

$$\hat{\mu}_{j\tau} = \hat{\gamma}_0 + \pi_1 \overline{VA}_{j\tau-1} + \pi_2 \overline{VA}_{j\tau-2} + \pi_3 \overline{VA}_{j\tau-3} + \pi_1 \overline{VA}_{j\tau+1} + \pi_2 \overline{VA}_{j\tau+2} + \pi_3 \overline{VA}_{j\tau+3} \qquad (8)$$

In equation (8), $\overline{VA}_{jt}$ is estimated value-added for teacher *j* in year *t*. Finally, we can construct the residualized student achievement measures as described above, $\hat{A}_{i\tau}$, and estimate the following regression:

$$\hat{A}_{i\tau} = \alpha_\tau + \theta \hat{\mu}'_{j\tau} + \xi_{i\tau} \qquad (9)$$

The predictor of interest in equation (9), $\hat{\mu}_{j\tau}$, is interpreted as the best linear predictor of $\hat{A}_{i\tau}$ under the assumption that stationarity of teacher value added is upheld through the transition year. This assumption is built into the regression by construction because $\hat{\mu}_{j\tau}$ depends on prediction coefficients (λ_1, λ_2, and λ_3) that represent correlations in value added over time obtained without using any data from transition years. Therefore, under the null hypothesis, our estimate of θ should be indistinguishable from one, just like our estimate of λ in equation (7), albeit less precisely estimated

due to efficiency costs associated with the restrictions we impose to partition off the data to estimate equations 8 and 9, as described above. Because we are concerned about how transition measures compare to stable measures, in practice we test whether our estimate of θ in the transition period is statistically distinguishable from our estimate of λ from the stable period, with the practical importance of the violation indicated by how far θ is from λ. In the extreme case, an estimate of θ = 0 would indicate that teacher performance in stable years has no predictive power over performance in transition years.[30]

IV. Results

As outlined in the analytics section above, we first present a descriptive overview of stability in estimates of teacher performance by measuring year-to-year correlations of value added in subsection A, top and bottom decile persistence in subsection B, and average teacher rankings by classroom type in subsection C. We then use regression analysis to formally measure changes in teacher percentile rankings in transition years in subsection D. Finally, in subsection E, we compare student scores in transition years to forecasted scores based on the performance of their teachers in stable years to measure the extent to which teacher performance in stable years can accurately predict performance in transition years.

A. Correlation of Value Added Across Test and Curriculum Changes

[30] The tests described in this section are not suited to determine whether stationarity is driven by teacher quality or some other biasing factor like persistent sorting. Chetty et al. (2014) address this concern with a teacher-switching quasi-experiment, where changes in student test scores at the grade-subject-school-year level are compared to changes in forecasted teacher value added in the same cell. However, the number of years that would have to be discarded to perform an analogous test for transition years make performing a high-powered test infeasible. To illustrate, consider a hypothetical scenario where five years of data are available, with the first being a transition year. To perform the validity test, we would first use stable years 2-5 to estimate the correlation of value added 1 year, 2 years, and 3 years apart. For example, the 3 year apart correlation would be the correlation of year 2 and year 5. We would then use these estimated correlations to forecast value added in years 2-4 using the year 1 transitional value added. For example, the estimated 3 year apart correlation would be used to forecast year 4 value added using the transition year: year 1. Differencing the data would then result in two observations from the original five years of data: the difference between year 2 and year 3, and the difference between year 3 and year 4. Thus, five years of data would have been used to create two years of grade-subject-school-year data for use in the test.

Before focusing on our main findings, a few ancillary results are worth mentioning. The magnitudes of the adjacent-year value added correlations, displayed in Figures 1a-1g with vertical bars denoting initial years of new assessments, are consistent with what has been found elsewhere in the literature (e.g., Chetty et al., 2014; Goldhaber & Hansen, 2013; Kane & Staiger, 2011; McCaffrey et al., 2009), although there is some cross-state variation. Also consistent with the literature, the adjacent-year correlations are higher in math than in reading in most years and most states. However, there are clear differences among the states in the adjacent-year correlations between the two subjects. In North Carolina, for instance, the math correlation exceeds the reading correlation by about 0.20, whereas in Massachusetts and Washington, the differential in adjacent year correlations between math and reading is much smaller.[31]

In our main results, the first important pattern is that adjacent-year correlations are not constant across stable regimes. This is most readily apparent in North Carolina, the state with the longest panel of available data, where there is considerable year-to-year fluctuation in adjacent-year correlations even when the assessment regime is constant. For example, there are drops in correlations in math in 2005 and reading in 2006. For the most part, correlations in the transition year are broadly similar to correlations in stable years. Thus, in most instances of new assessments, there is no evidence that the year-to-year correlation of teacher value-added falls during transition years in these states. The exceptions are Kentucky and Massachusetts, where the correlation between value added in the year before transition and value added in the transition year drops noticeably in math and even more so in

[31] In a separate analysis, not shown, we find that much of the difference between correlations in math and reading in North Carolina is due to the greater measurement error in the reading test as is apparent from the much smaller differences across subjects in both states once the a adjacent year correlations have been adjusted for measurement error (see Goldhaber and Hansen, 2013, for a description of the adjustment process). That adjusting the correlations makes a bigger difference for reading than math has been found before in the North Carolina data (e.g., Goldhaber and Hansen, 2013).

reading.[32] Below, we investigate the implications of the reduced correlations in these two states for identifying particularly high- and low-value-added teachers.[33,34]

B. Exploring the Tails of the Distribution

Although the correlations presented above are informative about the general relationship of value added one year to the next, the tails of the distribution are especially relevant for policy. Teacher evaluation systems that have been implemented in practice thus far have focused primarily on identifying teachers in the tails of the quality distribution for high-stakes intervention (e.g., Washington DC IMPACT, the Tennessee Educator Evaluation System). A primary objection to evaluation based on CCSS-aligned assessments is that teachers are not prepared for the new tests, so it is not fair to use them in personnel evaluations without allowing time for teachers to adapt, especially because CCSS is designed to be more rigorous than previous statewide standards.[35] With this in mind, in this section we measure cross-year changes in the likelihood that teachers remain in the top and bottom deciles of teacher value added during stable and transition-year regimes. Specifically, we take teachers whose value added placed them in the top/bottom 10 percent in year *t-1* and measure the share who remain in the top/bottom

10 percent in year *t* for teachers who were observed in both *t-1* and *t* under each regime type.

[32] In results not shown, we measure correlations across regime types by measuring two-year apart correlations in value added. On the whole, they are largely similar to what the one year apart correlations.

[33] As noted above, one factor that could potentially contribute to the effect of transition years on the year-to-year correlation of value-added is a change in the predictive power of the lagged-test-score controls over current-year test scores for students in transition-year models. Less predictive lagged-score controls will result in noisier models, and thus noisier estimates of teacher value-added. However, for the transitions we study, the predictive power of lagged achievement is essentially unchanged during transition years relative to nontransition years. The major exception is in Kentucky, where beginning in 2013 the predictive power of pretest scores rises substantially.

[34] The drop in MCAS correlations in Massachusetts middle school math and reading is puzzling because for these districts, there was no standards or assessment change. Correspondence with the state's Department of Education provided no ready explanation. However, despite this dip, in results available from authors, we find that the maximum forecast deviation in MCAS districts is on the order of 10 percent (compared to 40 percent in reading for the PARCC districts).

[35] Chang, Kenneth. "With Common Core, Fewer Topics but Covered More Rigorously." *The New York Times*. September 2, 2013.

Results for these transition likelihoods are presented in Table 3. Consistent with previous work (e.g., Goldhaber & Hansen, 2013), rankings in stable regimes tend to be more volatile at the bottom of the distribution, reflected here by a general pattern of smaller persistence shares in the bottom decile than the top. This continues to be true through the transitions, and the patterns in Table 3 largely mirror what we show in the correlations presented above: each stable regime is associated with likelihoods of being consistently identified in the top and bottom deciles, and the likelihoods during transition periods are similar to those in the surrounding stable regimes, with the continued exceptions of Kentucky and Massachusetts, especially in reading. For example, among math teachers in Washington, 38 and 36 percent of teachers who were in the top decile in one year remained in the top decile the following year during the first and second stable regimes, respectively. In the transition year, the share was 38 percent.

Of the 32 transition * subject * decile instances where we observe a transition period surrounded by two stable regimes (i.e., in all transitions but the second transition in Massachusetts, where we do not observe a stable regime after this transition), in only five cases does classification consistency in transition periods fall more than one percentage point below the range given by the two surrounding stable regimes (Kentucky top decile math, Kentucky bottom decile reading, NYC elementary top decile reading, and both the top and bottom deciles for reading in Washington). When breaking down these deviations by subject, this corresponds to 1 case in 16 for math and 4 cases in 16 for reading. In only one case out of 32—the bottom decile for reading in Kentucky—is the difference between the transition value and the surrounding stable values statistically significant; this case is also the only one where the transition value is more than four percentage points outside the range of the surrounding stable values. In Kentucky, the share of reading teachers in the bottom decile in one year who remained there the following year fell from 29 percent before the transition to 18 percent in the transition period before rebounding to 26 percent in the second stable period. In some instances we

even observe higher classification consistency in transition years (e.g., the top decile for North Carolina's first math transition), although this is rare.

In addition to the transition in Kentucky, the other case in which we see large changes in transition years is the second transition in Massachusetts. The largest change in math is in this second transition in middle school for Massachusetts, where the share of teachers remaining in the top decile fell by 10 percentage points relative to before the transition. However, because we do not observe a stable regime following this transition, we cannot know whether the drop in classification consistency is due to transition year volatility or simply reflecting what consistency will be in the following stable regime.

C. Teacher Rankings By Classroom Type

A number of factors associated with any particular regime shift can influence how individual teachers are affected. For example, a new assessment might differ in its targeting and/or ceiling properties (Koedel & Betts, 2010), and this might disproportionately influence the rankings of teachers who teach certain types of students. This may be especially relevant for CCSS transitions as in many cases CCSS-aligned tests are more rigorous than the tests they replace (e.g., Lestch et al., 2013). Another potential source of cross-teacher variation may be teachers in disadvantaged settings facing additional challenges in adapting to new standards and assessments due to time and resource constraints.

In Table 4 we examine how teachers are ranked based on value-added in stable and transition years for three types of classrooms using definitions from Goldhaber et al. (2013): advantaged classrooms, which fall into the top quintile of prior year achievement (averaged across math and reading) and the bottom quintile of percent FRL for a given year; average classrooms, in the middle quintile of prior year achievement and percent FRL; and disadvantaged classrooms, defined to be the lowest scoring on prior year achievement and highest quintile of FRL students.[36] Our approach is

[36] The FRL measure in Kentucky varies wildly over time so we use only classroom achievement to define

straightforward: we estimate teacher value-added for each year of our data panel in each state for all teachers, and then we report the average percentile rank of teachers by classroom type from stable and transition years. Per the above, a transition year is defined as a year in which the pretest and posttest do not match. Because the results in Table 4 are based on single-year value-added estimates (not correlations spanning multiple years), transition years are always defined as the initial year in which a new curriculum and assessment regime is implemented.

Here, there is little evidence that teachers in disadvantaged classrooms disproportionately struggle in transition years. If anything, the average value-added percentile ranking of teachers in disadvantaged classrooms is somewhat higher in many of the transitions we observe. Thus, we find no evidence that teachers placed in disadvantaged classrooms fare worse on value-added measures in transition years relative to stable years.

D. Regressions Predicting Change in Volatility of Teacher Ranking

In addition to year-to-year correlations of teacher value added and decile persistence, another way to measure volatility is simply to measure the change in a teacher's percentile ranking from one year to the next. We formally test whether transition years are associated with greater volatility by this measure by regressing the absolute value of the change in each teacher's percentile ranks from year *t-1* to year *t* on classroom characteristics and whether value added in either year was a transitional estimate. If switching to a new assessment were associated with increased volatility in teacher rankings, one would expect a positive and significant coefficient on the transition term. Results are shown in Table 5. Odd-numbered columns show a base specification with no interaction terms between transition years and other explanatory factors. Teachers in classrooms with high percentages of FRL and Black students tend to have slightly more volatile rankings in most sites.[37] In all instances, teacher rankings for reading

advantaged classrooms for that state.

[37] We do not control for FRL in Kentucky due to large discrepancies in the year-to-year share of students identified as FRL in Kentucky. In some years, the share of FRL students is very low, so it is not feasible to divide classrooms

are more volatile than for math, as evidenced by negative math coefficients. This result is likely driven by a lower signal-to-noise ratio in estimates of teacher value added in reading (Lefgren & Sims, 2012).

Most relevant to the current analysis, the coefficients on transition years in all states but Kentucky represent changes of less than one percentile point, and are even negative in some cases. Consistent with the correlations presented above, in states other than Kentucky, these results suggest that transition-year value added is not substantially more volatile than value added in any other year. On the other hand, Kentucky shows a moderate increase in volatility of about 10 percent relative to baseline volatility: an increase of 2.5 teacher percentile ranks relative to an average year-to-year change of about 25. Although many of the transition coefficients are statistically significant, no other site shows an increase in volatility of more than 5 percent, and this is well within the year-to-year variation we observe in a stable standards and assessment regime. The addition of the interaction terms in even-numbered columns provides no new insights aside from evidence that the volatility in teacher rankings in transition years in Kentucky is largely driven by reading and that there is substantial variability in Massachusetts in middle school as well, which is not surprising given the year-to-year correlations presented above. Overall, results from Table 5 indicate that transition years are not associated with meaningful changes in the volatility of teacher rankings in most cases.

E. Estimates of Forecast Instability

Table 6 displays the results of the forecasting exercise described by equation (9). The key feature of this test is that we do not use any data from transition years in constructing the value added forecasts. We begin by showing results from equation (7), estimated during stable periods only. Results are shown in column 1 and the forecasting coefficients are close to unity by construction. The exception is in Massachusetts, where we observe deviations from unity in middle grades, likely because the limited

into groups based on percent FRL.

number of stable years in the data. In North Carolina, the coefficient is almost exactly 1, while in Washington and Kentucky, we estimate a coefficient of 0.97 during the stable period, which is in some cases (Washington) statistically significant, but still implies that forecasted and actual test scores generally track each other tightly in stable periods. In NYC, the coefficients are about 0.02 and 0.04 units away from 1 in elementary and middle school, respectively.

Next, we use the auto-correlation of teacher value added that was estimated without using any data from transition years to forecast teacher effectiveness in transition years as a test for whether teacher value-added maintains the property of stationarity through transition years. Results for all transition years in both subjects are pooled together and shown in column 2.[38] In North Carolina and Washington, we cannot reject that teacher performance, as measured by value-added, is stationary though the transition year. However, in each of the other states, we reject the null hypothesis of equality between the stable (column 1) and transition (column 2) coefficients and thus conclude that stationarity is not maintained through the transition.

In columns (3) – (6), we disaggregate across subjects and transitions to gain further insight. The disaggregation shows that the divergences are driven by violations to stationarity for the reading transitions, where our coefficients are far from one. In math, most of the informational content of value-added is generally maintained throughout the transitions. Estimates of forecast deviations range from 1.6 percent to 8.9 percent in elementary school and 8.3 to 12.2 percent in middle school. In reading, estimates of forecast instability are often much larger, ranging from 2.0 to 41.4 percent in elementary school and 7.3 to 42.8 percent in middle school. Cases of large forecast deviations, however, are relatively infrequent: across the 20 transition * subject * school level cells, only three have forecast deviations greater than 16 percent, all in reading. And while teacher performance in transition years

[38] When plotting the relationship between test score residuals and forecasted value added in transition years, the estimated relationship appears to hold at all points in the test score distribution.

often cannot be forecasted with the same accuracy as in stable years, each of the transition year coefficients is very far from zero, indicating that teacher performance in stable years does provide useful information about performance in transition years.[39]

V. Discussion

In general, we find that the largest cases of volatility in teacher rankings (Table 5) and departures from forecast stability (Table 6) occur in reading transitions. One potential explanation for this pattern is that the typical scope for revising reading standards/tests is larger than for math, so new reading tests could measure different domains to a larger extent than new math tests. We find evidence that this may be the case when benchmarking revised assessments in Kentucky against a nationally administered test that did not change during the transition year. In Figures 2a and 2b, we display the correlation between a student's score on 8th grade Kentucky end-of-grade test, which was changed in 2012, and the ACT-administered EXPLORE test, which was not. While 8th graders are not included in this study, this exercise provides us with a chance to examine changes in Kentucky EOG tests relative to a stable test. As shown in Figures 2a and 2b, although the math EOG had roughly the same correlation with EXPLORE through the transition, this was not the case in reading. We take this as suggestive evidence that the reading test in Kentucky changed more than the math test.

[39] Another way to conceptualize whether value added estimates in transition years provide useful information is to consider the prediction of teacher quality in the year following a transition year. In results omitted for brevity, we find that including the transition year improves forecasting accuracy relative to excluding it, even during the reading transitions where our forecasts going into the transition are lowest. There are two reasons for this. First, including the transition year measure improves precision by incorporating additional data that the previous tables show to be informative. Second, in the handful of cases where there are large differences across tests, using the transition year improves the forecast by adding information from the new test and thus downweighting the contribution of the old test to the forecast of post-transition value-added. To illustrate, consider the exercise of forecasting reading value added in 2013 in Kentucky, which had its transition in 2012. When not using 2012, the coefficient from the regression of student scores on forecasted value added is .579. Adding 2012 to the forecast increases the coefficient to .779, meaning that incorporating the transition year results in a better forecast for 2013 than excluding it, reducing the forecast error by about half. On the other hand, in states like North Carolina where the transition is very stable, incorporating the transition year generally improves precision but the forecasts are accurate whether or not the transition year is included because teacher performance across tests is stable.

In multiple cases, we see the year-by-year correlation of value added drop in math in transition years without a corresponding drop in forecast stability (e.g., Kentucky and middle grades in Massachusetts, where the transition year estimates in Table 6 are similar to the stable year estimates). There are three factors at play that can explain the apparent puzzle of why math generally shows forecast stability in transition years even when the correlation drops. First, using Kentucky as an illustration, the correlation between value added in 2011 and value added in 2012 does not fall nearly as far in math as in reading: math still carries a relatively high signal in the transition year. Second, in contrast to math, the reading correlation never rebounds to the prechange levels. This violates the assumption of stationarity since these one-year-apart correlations are not constant over time, even in stable regimes: the correlation is distinct in the two separate stable regimes. This failure of the correlations to rebound to prechange levels is also visible to a lesser extent in Washington's reading transition (Figure 1g) and likely explains the divergence from 1 in column 5 of Table 6 for Washington. Finally, for math in Kentucky, even though 2011 and 2012 have a lower correlation (relative to other one-year-apart correlations), 2011 is one of five years used to construct the forecast of value added in 2012: 2009-2011 and 2013-2014. In particular, in results not shown, we find that value added in 2013 and 2014 are still highly correlated with 2012, so the forecast can still predict student achievement successfully. In contrast, in reading, none of the posttransition year-to-year correlations are as high as the pretransition correlations.

Although many possible factors may explain the increased volatility in reading value added during the Massachusetts and Kentucky transitions, we can rule out four explanations. First, assessment changes may lead to differences in how well students at different parts of the test score distribution are targeted. For example, a shift to a more rigorous test may lead to improvements at targeting high-ability students. We characterize the difference in the distribution between old and new tests by measuring the density divergence (see Frölich, 2004), which is meant to provide a broad indication of the degree of

change in test content and targeting. Density divergences in each state are generally similar in math and reading, and thus density divergence cannot explain why reading tends to have a higher degree of forecast instability. Second, while density divergence provides a general measure of distributional shifts, it may be at the tails of the distribution where test targeting is most relevant (e.g., Koedel & Betts, 2010). Thus, motivated in part by the distribution of Kentucky's older test (see Figure A1), we measure the skewness of pretransition tests at each of our research sites. We find that the skewness in Kentucky does not differ substantially from other states where we do not find forecast instability, and, as noted earlier, the skewness in Kentucky is well below the threshold that Koedel and Betts (2010) find distorts measures of teacher performance. Third, we examine the predictive power of prior test scores. In all cases aside from Kentucky, the coefficient on lagged test scores in transition years is similar to stable years in each subject. And the change in pretest predictive power cannot explain the volatility we see in Kentucky because the predictive power of same-subject lagged scores rises dramatically in both math and reading under the new regime, whereas only in reading do we observe forecast instability. Finally, when adjusting year-to-year correlations to remove sampling error (described in Appendix B), the transitions with large unadjusted drops in correlations continue to have large drops after the adjustment, ruling out increased sampling error in transition years as an explanation.

The states that we study provide suggestive evidence about the relative importance of standards versus assessments. If revisions to standards were driving instability in teacher rankings due to changes in the content students are expected to learn, one would expect to see stable teacher rankings in states that adopted new tests while maintaining stable standards. However, two of the three largest instances of forecast instability (Massachusetts when adopting PARCC and Washington) occurred when a state administered a new reading test without any change in standards. Thus, properties of the tests themselves, rather than changes in standards, may be driving the instability we find in teacher performance measures.

VI. Policy Implications and Conclusions

Although it is not possible to predict with certainty how the rollout of CCSS-aligned standards and assessments will affect the quality of the information contained in estimates of teacher value added, our investigation of prior standards and assessment changes—including some changes involving the CCSS—provides insights to help guide decisions about teacher evaluations during the transition. The evidence presented above indicates that previous standard and assessment changes *in math* have had minimal effects on the stability of estimated teacher value added and teacher rankings at the tails of the distribution. In addition, in most cases we find that teacher performance in stable regimes can forecast student test scores in transition years with a high degree of accuracy. The findings in math are consistent with a growing body of evidence showing that value added is a meaningful and persistent measure of teacher quality (Koedel et al., 2015), and reveal that this can be the case even during standard and assessment transition years.

In reading, on the other hand, the relationship between past teacher performance and performance in transition years is weaker— substantially so in some cases. One explanation is that the content of tested material changed more dramatically in reading than math for the transitions we study in ways that are difficult to quantify. It may also be that our reading results are influenced by the lower reliability of teacher value added in reading more generally (also see Goldhaber & Hansen, 2013; Lefgren & Sims, 2012). However, although during some transitions the informational content of reading value-added estimates is degraded, we also note that in no instance do the measures cease to be informative about teacher performance. Even in the most volatile transition years, an increase of one standard deviation in forecasted student test scores—as forecasted by teacher performance in stable years—is associated with an approximately 0.60 standard deviation increase in observed scores. Thus, we show that a moratorium on value added scores in transition years would discard information that is clearly predictive of teacher performance. In doing so, it would implicitly increase the weight given to nontest

measures of teacher performance in transition years, and there is little evidence regarding how these measures perform during transitions.

References

Bacher-Hicks, A., Kane, T.J., & Staiger, D.O. (2014). Validating teacher effect estimates using changes in teacher assignments in Los Angeles. *NBER working paper No. 20657.*

Ballou, D. (2009). Test Scaling and Value-Added Measurement. *Education Finance and Policy*, *4*, 351-383.

Bazemore, Mildred, Pam Van Dyk, Laura Kramer, Amber Yelton and Robert Brown (2006). North Carolina Mathematics Tests: Technical Report. *The Office of Curriculum and School Reform Services*, North Carolina Public Schools.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the Impacts of Teachers I*:* Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review* 104(9): 2593-2632.

Corcoran, S. P., Jennings, J. L., & Beveridge, A. A. (2011). Teacher effectiveness on high-and low-stakes tests.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. J. (2014). The Sensitivity of Value-added Estimates to Specification Adjustments: Evidence from School-and Teacher-level Models in Missouri. *Statistics and Public Policy*, *1*(1), 19-27.

Frölich, M. (2004). Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators. *The Review of Economics and Statistics* 86 (1): 77-90

Fox, J. C. (2015, November 17). Education board votes to adopt hybrid MCAS-PARCC test. *The Boston Globe*. Retrieved from https://www.bostonglobe.com/metro/2015/11/17/state-education-board-vote-whether-replace-mcas/aex1nGyBYZW2sucEW2o82L/story.html

Glazerman, S., A. Protik, B. Teh, J. Bruch, J. Max. (2013). Transfer Incentives for High- Performing Teachers: Final Results from a Multisite Experiment (NCEE 2014-4003). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Glazerman, S., & Protik, A. (2015). Validating value-added measures of teacher performance. *Unpublished manuscript*.

Goldhaber, Dan, and Hansen, Michael. (2013). Is it Just a Bad Class? Assessing the Long-term Stability of Estimated Teacher Performance. *Economica*, Vol 80(319), pp 589–612.

Goldhaber, Dan, Gabele, Brian, and Joe Walch (2013). Does the Model Matter? Exploring the Relationship Between Different Student Achievement-based Teacher Assessments. *Statistics and Public Policy*. Vol 1(1), pp. 28–39.

Guarino, C. M., Maxfield, M., Reckase, M. D., Thompson, P. N., & Wooldridge, J. M. (2015a). An Evaluation of Empirical Bayes's Estimation of Value-Added Teacher Performance Measures. *Journal of Educational and Behavioral Statistics*, 40(2), 190-222.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015b). Can Value-Added Measures of Teacher Performance Be Trusted? *Education Finance and Policy*, *10*(1), 117–156.

Helms, Ann Doss. (2013). "New N.C. exams test teachers – in more ways than one." *Charlotte Observer*.

Innes, Richard. Federal tests show Kentucky's test scoring inflated again in 2009. *Bluegrass Institute*. October 15, 2009. http://www.bipps.org/federal-tests-show-kentucky%E2%80%99s-test-scoring-inflated-again-in-2009/

Kane, T. J., & Staiger, D. O. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation (No. w14607). *National Bureau of Economic Research*.

Kane, T. J., & Staiger, D. O. (2011). Initial Findings from the Measures of Effective Teaching Project. *Bill and Melinda Gates Foundation.*

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. *Seattle, WA: Bill and Melinda Gates Foundation*.

Koedel, Cory and Julian R. Betts (2010). Value-Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation. *Education Finance and Policy* 5(1): 54-81.

Koedel, Cory, Leatherman, Rebecca, & Parsons, Eric (2012). Test Measurement Error and Inference from Value-Added Models. *The B.E. Journal of Economic Analysis & Policy 12*(1).

Koedel, C., & Li, J. (2016). The Efficiency Implications of Using Proportional Evaluations to Shape the Teaching Workforce. *Contemporary Economic Policy*, 34(1), 47-62.

Koedel, Cory, Kata Mihaly & Jonah Rockoff (2015). Value-Added Modeling: A Review. *Economics of Education Review* 47: 180-195.

Koretz, D., Marcus Waldman, Carol Yu, Meredith Langi, & Aaron Orzech (2014). Using the Introduction of a New Test to Investigate the Distribution of Score Inflation. *Harvard College*.

Lefgren, Lars & David Sims (2012). Using Subject Specific Test Scores Efficiently to Predict Teacher Value-Added. *Educational Evaluation and Policy Analysis* 34(1): 109-121.

Lestch, C., B. Chapman, & J. Fermino. (2013). City Students' Scores Take Dramatic Plunge on New Standardized Tests. *NY Daily News*.

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), 47-68.

Lockwood, J.R. and Daniel F. McCaffrey (2014). Correcting for Test Score Measurement Error in ANCOVA Models for Estimating Treatment Effects. *Journal of Educational and Behavioral Statistics.* 39(1), 22-52.

Massachusetts Department of Elementary and Secondary Education. (2004a). *Supplement to the Massachusetts English Language Arts Curriculum Framework*. Massachusetts Department of Elementary and Secondary Education.

Massachusetts Department of Elementary and Secondary Education. (2004b). *Supplement to the Massachusetts Mathematics Curriculum Framework*. Massachusetts Department of Elementary and Secondary Education.

Massachusetts Department of Elementary and Secondary Education. (2011b). *Massachusetts Curriculum Framework for Mathematics*. Massachusetts Department of Elementary and Secondary Education.

Massachusetts Department of Elementary and Secondary Education. (2011a). *Massachusetts Curriculum Framework for English Language Arts and Literacy*. Massachusetts Department of Elementary and Secondary Education.

Massachusetts Department of Elementary and Secondary Education. (2014a, March 18). Assessment Transition Plans - Massachusetts Comprehensive Assessment System. Retrieved November 24, 2015, from http://www.doe.mass.edu/mcas/transition/?section=math3-8

Massachusetts Department of Elementary and Secondary Education. (2014b, March 18). Assessment Transition Plans: Assessment Transition from 2001/2004 MA English Language Arts (ELA) Framework to 2011 MA Curriculum Framework for English Language Arts & Literacy - Massachusetts Comprehensive Assessment System. Retrieved November 24, 2015, from http://www.doe.mass.edu/mcas/transition/?section=ela

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.

Office of the Superintendent of Public Instruction. (2004a). Mathematics K-10 Grade Level Expectations: A New Level of Specificity (No. 04-0006). Olympia, WA: *Office of the Superintendent of Public Instruction*.

Office of the Superintendent of Public Instruction. (2004b). Reading K-10 Grade Level Expectations: A New Level of Specificity (No. 04-0001). Olympia, WA: *Office of the Superintendent of Public Instruction*.

Office of the Superintendent of Public Instruction. (2008). Washington State K-12 Mathematics Learning Standards. Olympia, WA: *Office of the Superintendent of Public Instruction*.

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.

Stacy, B., Guarino, C. M., Reckase, M. D., & Wooldridge, J. (2013). Does the Precision and Stability of Value-added Estimates of Teacher Performance Depend on the Types of Students they Serve? (No. 7676). Bonn, Germany: *Institute for the Study of Labor*.

Stepner, M. (2013) vam.ado [Version 2.0.1]. Retrieved from http://obs.rc.fas.harvard.edu/chetty/vam.ado.

Ujifusa, A. (2015, June 30). A "Common-Core Math" Problem: How Many States Have Adopted the Standards? Retrieved November 24, 2015, from http://blogs.edweek.org/edweek/state_edwatch/2015/06/a_common_core_math_problem_how_many_states_have_adopted_the_standards.html?cmp=SOC-SHR-FB

Winters, M. A., & Cowen, J. M. (2013). Would a value-added system of retention improve the distribution of teacher quality? A Simulation of Alternative Policies. *Journal of Policy Analysis and Management*, 32(3), 634-654.

Xu, Z., Ozek, U., & Corritore, M. (2012). Portability of Teacher Effectiveness Across School Settings. *CALDER working paper WP77*.

Zernike, K. (2015, November 21). Massachusetts's Rejection of Common Core Test Signals Shift in U.S. *The New York Times*. Retrieved from http://www.nytimes.com/2015/11/22/us/rejecting-test-massachusetts-shifts-its-model.html

Tables

Table 1: *Assessment Changes During Study Period*

State	Transition	Implementation Year (Spring)	Accompanied by Standards Change?
Kentucky	1	2012	Yes
Massachusetts	1	2013	Yes
	2	2015 (selected districts)	No
New York City	1	2006	Math: yes; Reading: no
North Carolina	1	Math: 2001; Reading: 2003	Yes
	2	Math: 2006; Reading: 2008	Yes
Washington	1	2010	Math: yes; Reading: no

Table 2. *Description of Data*

State	Years	Grades	Demographic Information	Unique Teachers	Unique Students
Kentucky	2009-2014	4-5	Race, gender, FRL, ELL, special education	7,577	297,347
Massachusetts	2011-2015	4-8	Race, gender, FRL, ELL, special education	24,977	709,863
New York City	Math: 2000-2010; Reading: 2003 (gr 5) or 2004 (gr 4) - 2010	4-8	Race, gender, FRL, ELL, disability/special education	26,519	823,389
North Carolina	1997-2012	4-5	Race, gender, FRL, disabilities	28,207	1,214,113
Washington	2006-2013	4-5	Race, gender, FRL, ELL, gifted/disability status	10,036	447,375

Notes: Years and grades indicate for which teachers value added can be computed. Additional data are used to compute value added scores; in Kentucky, for example, scores from third graders and from the 2008 year are used for pretest scores.

Table 3. *Likelihood of Top and Bottom VA Decile Teachers in t-1 Remaining in That Decile in t*

Kentucky

	Math		Reading	
Regime	Top	Bottom	Top	Bottom
Stable 1	0.34	0.31	0.32	0.29
Transition	0.28	0.26	0.24	0.18**
Stable 2	0.32	0.26	0.24	0.26

Massachusetts

	Elementary				Middle			
	Math		Reading		Math		Reading	
Regime	Top	Bottom	Top	Bottom	Top	Bottom	Top	Bottom
Stable 1	0.35	0.35	0.53	0.30	0.53	0.36	0.42	0.49
Transition 1	0.38	0.31	0.44	0.32	0.49	0.38	0.47	0.42
Stable 2	0.35	0.29	0.38	0.30	0.49	0.33	0.36	0.35
Transition 2	0.39	0.32	0.29	0.25	0.39	0.26	0.23	0.27

New York City

	Elementary				Middle			
	Math		Reading		Math		Reading	
Regime	Top	Bottom	Top	Bottom	Top	Bottom	Top	Bottom
Stable 1	0.37	0.29	0.38	0.30	0.39	0.34	0.33	0.29
Transition	0.37	0.27	0.35	0.30	0.45	0.36	0.34	0.27
Stable 2	0.35	0.28	0.38	0.24	0.44	0.38	0.34	0.27

North Carolina

	Math		Reading	
Regime	Top	Bottom	Top	Bottom
Stable 1	0.38	0.36	0.25	0.25
Transition	0.40	0.32	0.23	0.22
Stable 2	0.33	0.28	0.22	0.20
Transition	0.34	0.33	0.24	0.20
Stable 3	0.35	0.28	0.21	0.20

Washington

	Math		Reading	
Regime	Top	Bottom	Top	Bottom
Stable 1	0.38	0.25	0.35	0.26
Transition	0.38	0.28	0.31	0.23
Stable 2	0.36	0.29	0.33	0.25

Notes: Top shows the share of teachers who were in the top decile in year *t-1* who remained in the top decile in year *t*, while bottom shows the share of teachers in the bottom decile who remained in the bottom. Significance stars indicate transition year value is significantly lower than both surrounding stable regimes at the 90% (*), 95% (**), and 99% (***) significance levels. The second Massachusetts transition does not have a following stable regime, so no inference is displayed.

Table 4. *Average Teacher Percentile Ranks by Classroom Type*

Kentucky

	Math			Reading		
Regime	Adv.	Average	Disadv.	Adv.	Average	Disadv.
Stable 1	55.8	48.9	44.1	56.8	48.9	42.8
Transition	59.1	47.3	44.7	54.8	47.6	50.4
Stable 2	56.5	50.4	45.2	56.3	50.1	43.2

Massachusetts

	Elementary						Middle					
	Math			Reading			Math			Reading		
Regime	Adv.	Avg.	Disadv.	Adv.	Avg.	Disadv.	Adv.	Avg.	Disadv.	Adv.	Avg.	Disadv.
Stable 1	63.6	43.4	44.6	62.7	46.1	39.5	61.9	50.4	34.8	66.5	45.7	26.7
Transition 1	57.3	46.0	50.5	63.4	48.2	39.3	58.9	50.4	38.1	64.5	50.6	25.5
Stable 2	53.8	50.5	51.8	60.6	50.6	40.7	59.9	50.6	37.3	60.9	52.1	30.3
Transition 2	57.0	47.2	49.4	60.0	47.5	42.3	62.7	43.8	39.8	56.8	42.0	35.0

New York City

	Elementary						Middle					
	Math			Reading			Math			Reading		
Regime	Adv.	Avg.	Disadv.	Adv.	Avg.	Disadv.	Adv.	Avg.	Disadv.	Adv.	Avg.	Disadv.
Stable 1	61.2	48.9	39.1	72.8	47.6	33.5	63.9	43.1	45.0	72.4	38.7	34.6
Transition	63.4	45.8	41.3	67.6	46.8	38.9	64.2	49.9	37.2	71.1	49.6	35.1
Stable 2	64.9	45.1	45.3	67.4	45.2	42.6	65.7	47.9	39.9	72.2	46.8	37.4

North Carolina

	Math			Reading		
Regime	Adv.	Average	Disadv.	Adv.	Average	Disadv.
Stable 1	55.6	50.5	48.7	59.0	49.7	45.0
Transition	60.0	42.4	47.2	57.4	47.3	46.3
Stable 2	58.3	44.5	47.9	58.2	49.2	45.8
Transition	58.1	44.6	51.3	58.4	49.7	45.4
Stable 3	55.5	47.2	49.1	58.0	45.1	44.2

Washington

	Math			Reading		
Regime	Adv.	Average	Disadv.	Adv.	Average	Disadv.
Stable 1	61.4	48.0	44.1	60.0	47.6	46.9
Transition	59.7	49.7	47.6	54.4	49.5	45.5
Stable 2	55.6	47.3	50.7	55.6	49.2	44.5

Notes: Advantaged ("adv.") is defined as the top quintile of average prior achievement and the bottom quintile of percent FRL, average ("avg.") is the middle quintile of each, and disadvantaged ("disadv.") is the bottom quintile of average prior achievement and top quartile of percent FRL. Significance stars indicate transition year value is significantly lower than both surrounding stable regimes at the 90% (*), 95% (**), and 99% (***) significance levels. The second Massachusetts transition does not have a following stable regime, so no inference is displayed.

Table 5. *Prediction of Absolute Value of the Change in Teacher Percentile Ranking Between Year t and t-1*

	North Carolina		Washington		Kentucky	
	(1)	(2)	(3)	(4)	(5)	(6)
%FRL	2.70***	1.97***	0.46	1.11		
	(0.42)	(0.49)	(0.65)	(0.78)		
%Black	0.81**	1.18***	0.64	0.18	-1.90**	-3.53***
	(0.39)	(0.45)	(1.47)	(1.80)	(0.96)	(1.13)
Premath	-0.93***	-1.32***	0.98*	0.54	-0.84	-1.47
	(0.33)	(0.38)	(0.53)	(0.64)	(0.80)	(1.01)
Preread	0.226	0.07	0.13	0.66	-0.89	-1.31
	(0.36)	(0.41)	(0.59)	(0.71)	(0.84)	(1.01)
Math	-4.54***	-4.09***	-2.20***	-2.21***	-2.01***	-1.60***
	(0.12)	(0.18)	(0.22)	(0.26)	(0.30)	(0.37)
Transition	-0.55***	0.57	0.33	2.06***	2.50***	1.67**
	(0.13)	(0.46)	(0.25)	(0.72)	(0.33)	(0.67)
Transition * Math		-1.55***		0.02		-1.02*
		(0.45)		(0.47)		(0.59)
Transition * %FRL		-0.95		-2.06		
		(0.79)		(1.29)		
Transition * %Black		0.62		1.39		3.82**
		(0.67)		(3.01)		(1.79)
Transition * premath		1.02		1.31		1.38
		(0.64)		(1.05)		(1.58)
Transition * prereading		-0.68		-1.62		1.13
		(0.69)		(1.24)		(1.66)
Constant	26.27***	25.82***	24.27***	23.74***	25.66***	25.16***
	(0.22)	(0.33)	(0.38)	(0.44)	(0.31)	(0.44)
Observations	113213	113213	31096	31096	17482	17482
R-squared	0.015	0.018	0.004	0.005	0.007	0.010

Notes: Outcome variable is the absolute value of the difference in percentile ranking between year *t* and year *t*-1, measured on a 100 point scale. In Kentucky, the percentage of minority students (black and Hispanic) is used in place of the percentage of black students due to small cell size.

Table 5 (cont.) *Prediction of Absolute Value of the Change in Teacher Percentile Ranking Between Year t and t-1*

	MA Elem		MA Middle		NYC Elem		NYC Middle	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
%FRL	0.80	0.60	5.10***	6.43***	0.15	0.395	0.07	0.20
	(0.81)	(1.51)	(1.07)	(1.58)	(0.38)	(0.47)	(0.54)	(0.66)
%Black	2.19*	0.23	2.04	2.32	0.59*	0.409	0.69	0.26
	(1.21)	(2.38)	(1.36)	(2.27)	(0.33)	(0.36)	(0.42)	(0.49)
Premath	-0.32	-0.74	1.20*	1.50	0.47	0.23	0.08	-0.27
	(0.73)	(1.42)	(0.73)	(1.07)	(0.43)	(0.48)	(0.57)	(0.64)
Preread	0.71	-0.01	0.11	-0.56	-1.17***	-1.00**	-1.94***	-1.78***
	(0.64)	(1.18)	(0.76)	(1.15)	(0.41)	(0.46)	(0.57)	(0.64)
Math	-0.17	0.21	-1.32***	-0.78	-0.16	-0.1	-4.58***	-4.86***
	(0.24)	(0.42)	(0.34)	(0.48)	(0.17)	(0.21)	(0.23)	(0.30)
Transition	0.65**	0.35	0.78***	5.20***	0.44**	-0.22	-0.66***	-2.34**
	(0.26)	(0.75)	(0.28)	(0.88)	(0.21)	(0.84)	(0.24)	(1.15)
Transition * Math		-0.51		-0.85		-0.62*		0.32
		(0.47)		(0.55)		(0.36)		(0.47)
Transition * %FRL		0.35		-2.46		-0.26		-0.08
		(1.62)		(1.79)		(0.85)		(1.17)
Transition * %Black		1.97		-0.98		-0.01		1.52*
		(2.60)		(2.54)		(0.76)		(0.86)
Transition * premath		0.58		-0.50		0.05		1.53
		(1.39)		(1.33)		(1.03)		(1.32)
Transition * prereading		0.87		0.83		-0.09		-0.83
		(1.28)		(1.38)		(0.98)		(1.31)
Constant	21.24***	20.75***	18.04***	16.63***	23.15***	24.25***	24.55***	25.69***
	(0.40)	(0.70)	(0.47)	(0.67)	(0.36)	(0.57)	(0.52)	(0.78)
Observations	25917	25917	17581	17581	57481	57481	32688	32688
R-squared	0.002	0.005	0.009	0.025	0.001	0.002	0.019	0.02

Notes: outcome variable is the absolute value of the difference in percentile ranking between year t and year t-1, measured on a 100 point scale. In Kentucky, the percentage of minority students (black and Hispanic) is used in place of the percentage of black students due to small cell size.

Table 6. *Out-of-Sample Forecasts of Transition Year Value Added*

	Stable (pooled)	Transition (pooled)	Math		Reading	
			Transition 1	Transition 2	Transition 1	Transition 2
	(1)	(2)	(3)	(4)	(5)	(6)
Kentucky						
	0.972	0.852	1.025		0.586	
	(0.018)	(0.045)	(0.059)		(0.057)	
p-value		0.01	0.39		<0.01	
Massachusetts elementary						
	1.022	0.928	0.970	0.969	0.980	0.653
	(0.014)	(0.016)	(0.023)	(0.048)	(0.020)	(0.043)
p-value		<0.01	0.05	0.29	0.09	<0.01
Massachusetts middle						
	1.059	0.993	1.083	0.915	1.073	0.572
	(0.013)	(0.018)	(0.026)	(0.053)	(0.025)	(0.061)
p-value		<0.01	0.39	<0.01	0.61	<0.01
New York City elementary						
	1.023	1.068	1.089		1.039	
	(0.008)	(0.021)	(0.025)		(0.028)	
p-value		0.05	0.01		0.59	
New York City middle						
	1.042	1.107	1.122		1.081	
	(0.010)	(0.022)	(0.025)		(0.040)	
p-value		<0.01	<0.01		0.35	
North Carolina						
	1.003	0.995	1.037	0.995	1.003	0.891
	(0.006)	(0.012)	(0.021)	(0.021)	(0.033)	(0.028)
p-value		0.55	0.10	0.71	0.99	<0.01
Washington						
	0.973	0.933	0.984		0.841	
	(0.011)	(0.022)	(0.027)		(0.028)	
p-value		0.12	0.69		<0.01	

Notes: Each coefficient is generated by a regression of residualized student test scores on forecasted student scores, with forecasts generated based on teacher performance out of sample. A coefficient of one indicates that forecasted student scores are an accurate predictor of actual scores. p-values are for test of coefficient against the stable coefficient in column 1.

Figures

Figure 1a: Adjacent-Year Correlations in Kentucky, Elementary

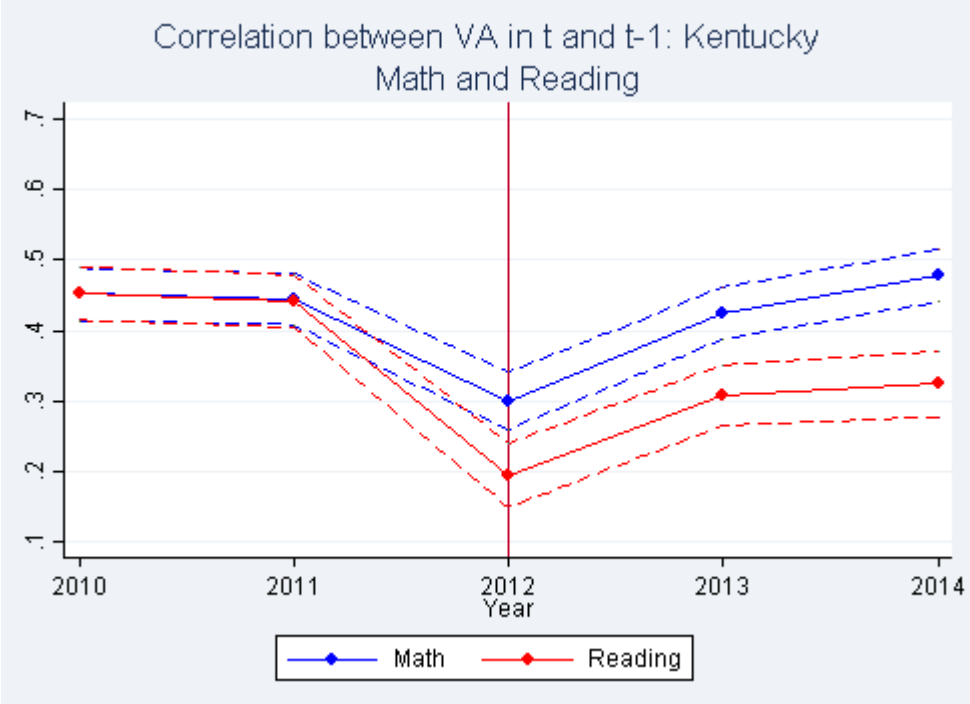


Figure 1b: Adjacent-Year Correlations in Massachusetts, Elementary

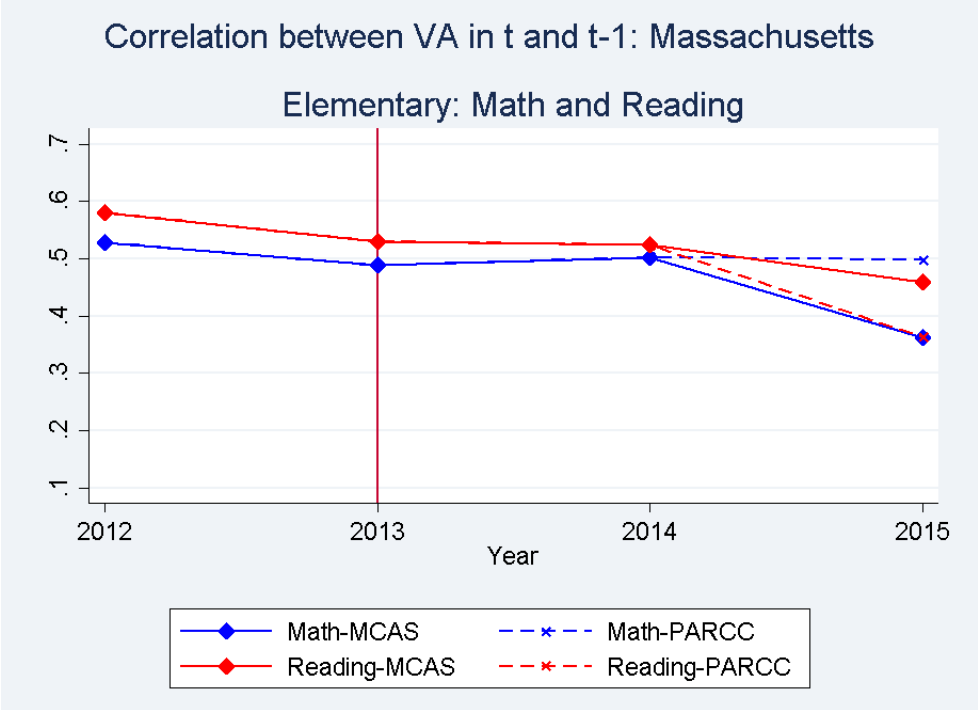


Figure 1c: Adjacent-Year Correlations in Massachusetts, Middle

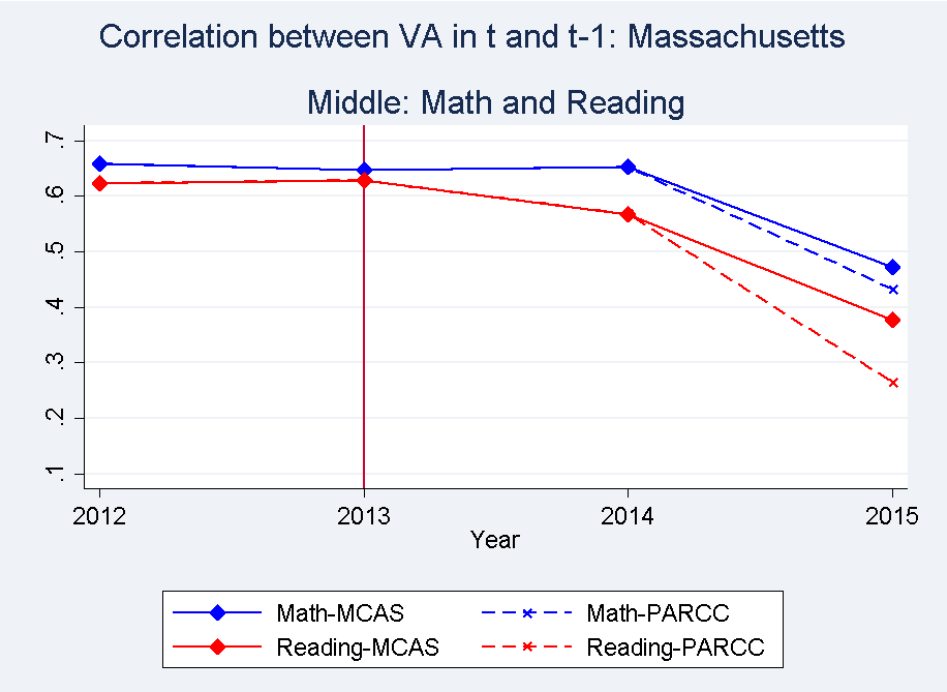


Figure 1d: Adjacent-Year Correlations in New York City, Elementary

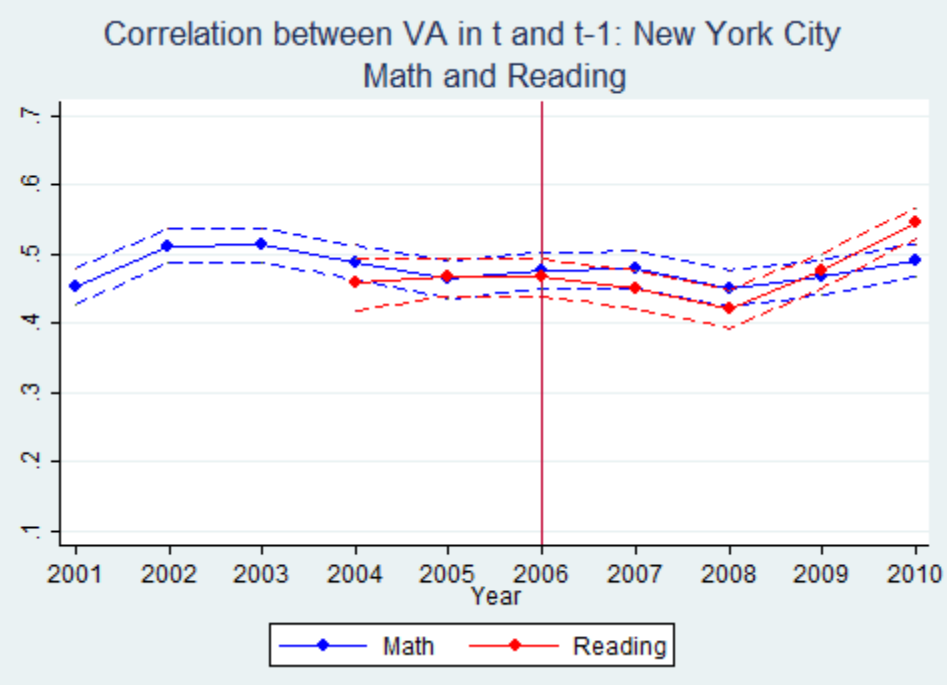


Figure 1e: Adjacent-Year Correlations in New York City, Middle

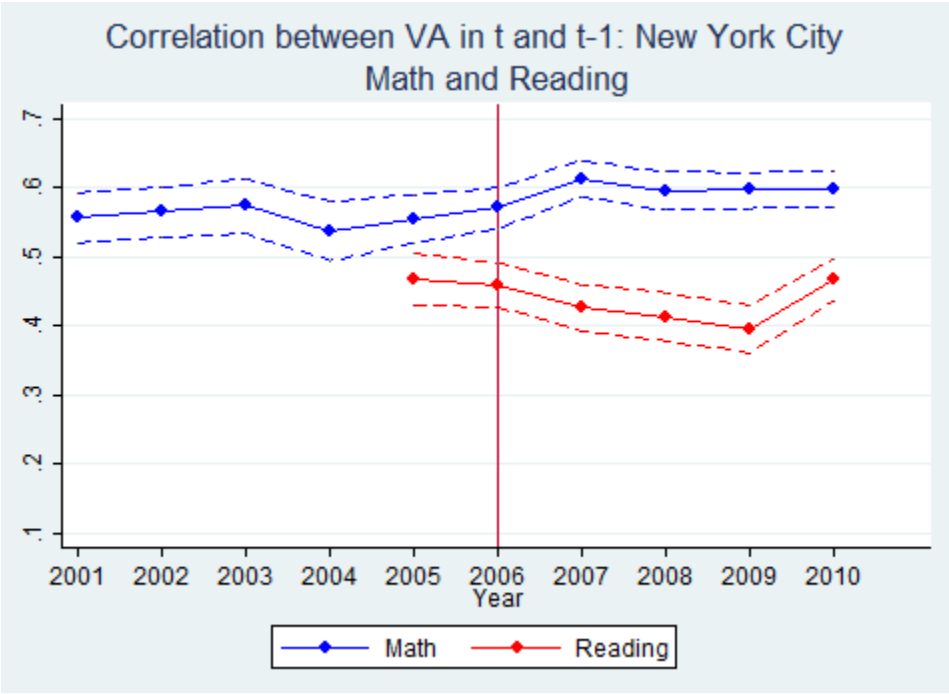


Figure 1f: Adjacent-Year Correlations in North Carolina, Elementary

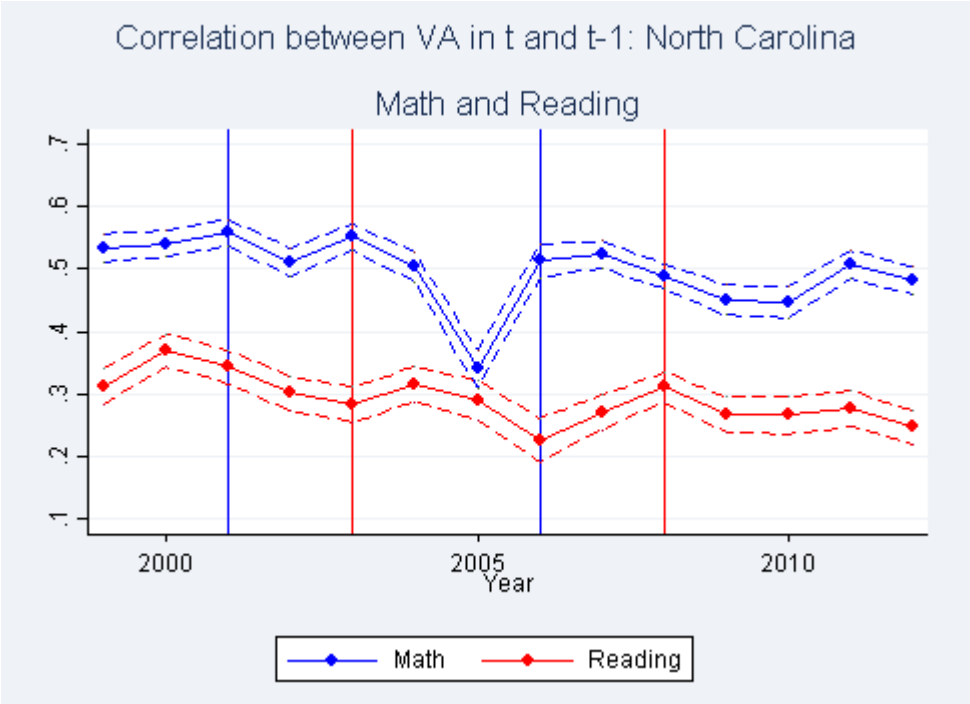
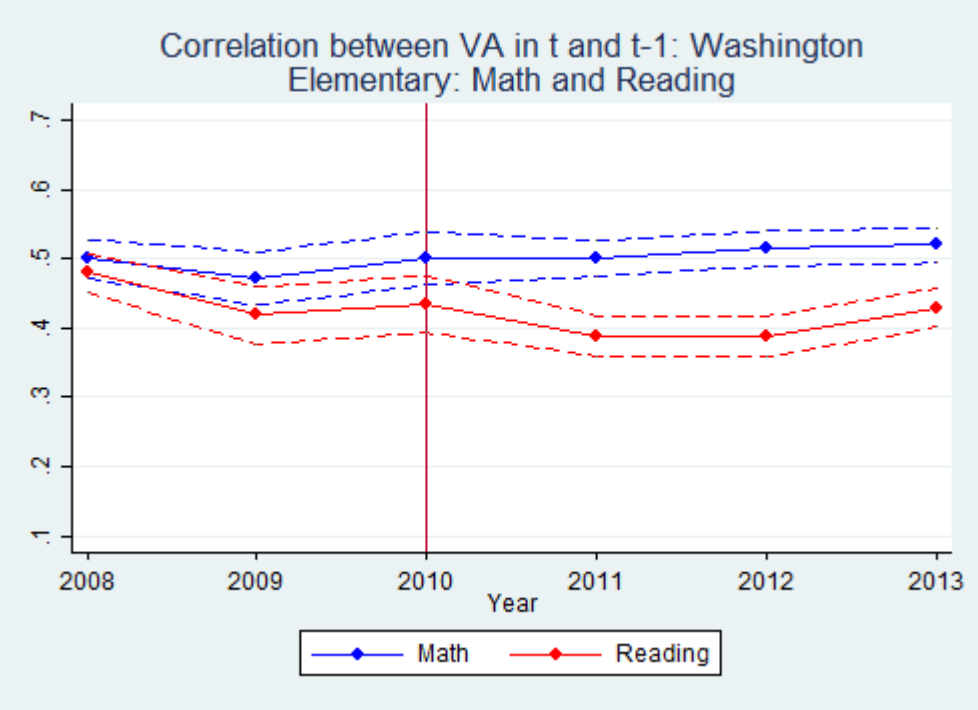


Figure 1g: Adjacent-Year Correlations in Washington, Elementary



Note: Vertical lines denote transition years. When math and reading transition years differ for a state, blue vertical lines denote math transitions and red lines denote reading transitions. Dashed lines represent standard errors. For Massachusetts, in 2015 districts had the choice of whether to remain administering the state's CCSS-aligned test (MCAS) or switch to PARCC; error bars in Massachusetts are suppressed for readability.

Figure 2a: Correlation Between EXPLORE and EOG in Kentucky, Math

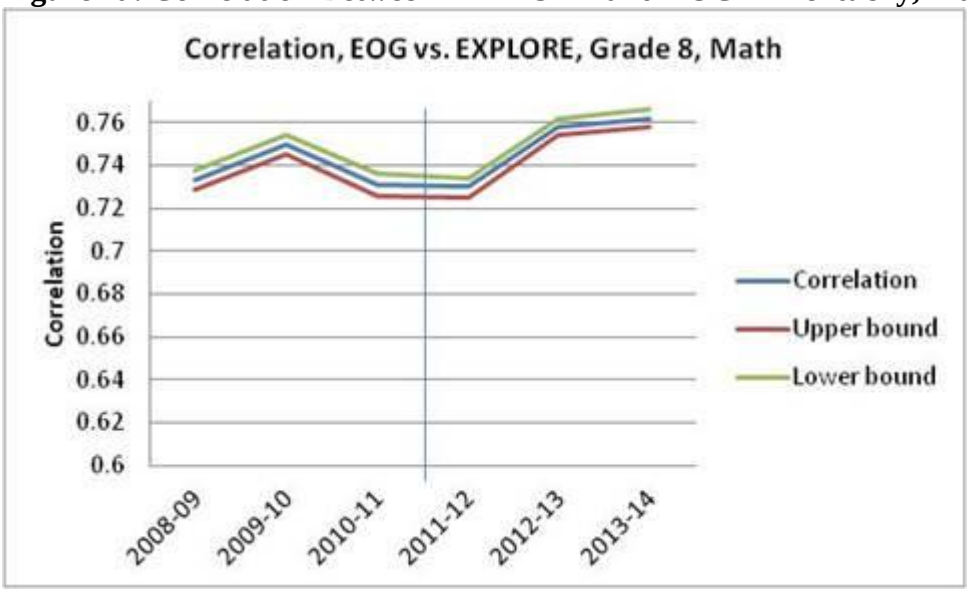
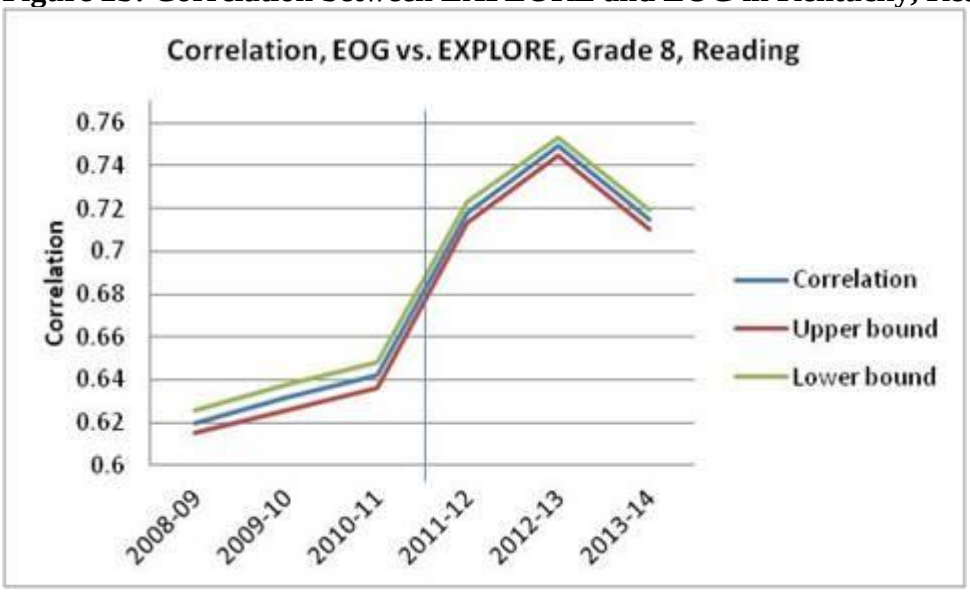


Figure 2b: Correlation between EXPLORE and EOG in Kentucky, Reading

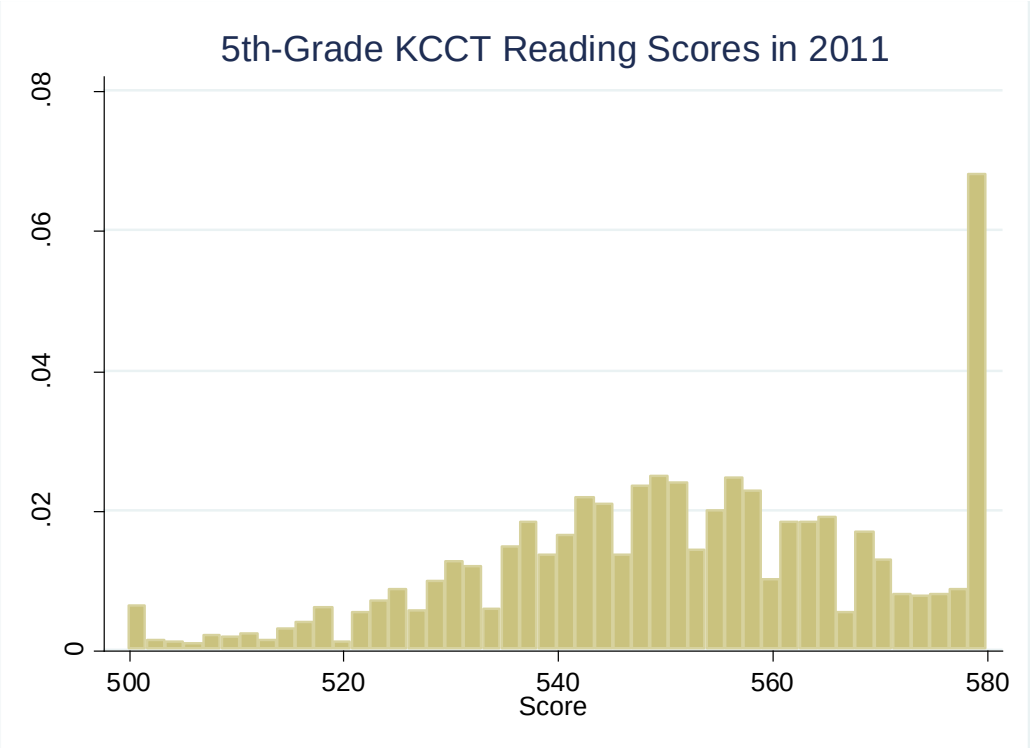


Appendix A – Additional State Information

Kentucky

In the years just before the testing regime change in Kentucky, the distribution of student test scores showed bunching to the right of the distribution. KCCT tests in third, fourth, and fifth grade were scaled to a range of 80 points, with as many as 10 percent of students receiving perfect scores in math and reading in each year. For example, the histogram below shows the distribution of fifth-grade reading scores in 2011:

Figure A1: Kentucky Test Distribution



As an alternate specification, we obtained results after following the two-step transformation used by Koretz et al. (2014) to normalize the distribution of test scores and reduce the potential for bias in value-added estimates arising from score inflation. First, we drop observations of students achieving the highest or lowest possible score on the 80-point scale, as well as observations of students achieving the highest or lowest possible score on the pretest in either subject. We then probit-transform the scores of the remaining students within year, grade, and subject. Results are similar when using these transformed scores, but we do not use them for the main results of the paper in part due to the dropping of students.

The forecast bias test in Table 6 requires each student to be linked to only one teacher in a given year and subject. Because many students in the sample were assigned to multiple classes and teachers within the same subject, we dropped classes that were not among the five most populated classes within year, grade, and subject, and then dropped any remaining students who were still observed with multiple teachers in a given subject and year. Finally, the data contain no identification of individual classrooms despite some teachers being assigned between 40 and 136 students within year, grade, and subject. To avoid limiting the sample any further, we treated each teacher-year-grade-subject observation as a classroom regardless of the number of students it contained, and dropped "classrooms" with 10 or fewer students.

The number of students identified as FRL-eligible jumped from about 12 percent in years prior to 2012 to about 60 percent in years 2012 and following. Because of concerns about the reliability of the variable, we do not control for FRL in Kentucky. Results are generally similar when adding FRL controls. For example, the estimate of forecast instability in Table 6 increases from 41.4 percent to 44.2 percent when adding FRL controls. The exception is in Table 5, where the specification with transition interaction terms yields negative Transition coefficients. This is likely due to the jump in measured FRL coinciding with the transition year, causing the specification where transition is interacted with FRL to produce anomalous results.

Massachusetts

As noted above, any students linked to multiple teachers were dropped. In math, this meant dropping 7 percent of students in elementary grades and 24 percent of students in middle grades. In reading, we dropped 13 percent of students in elementary grades and 41 percent of students in middle grades.

Among districts that administered the PARCC test in 2015, approximately one-third administered the test on paper, about half of districts administered the test online, and the remaining districts used a combination of the two test modes. Students taking the paper test in 2015 consistently scored better than students who took the test online. In theory, this result could be driven either by characteristics of the paper test or by students who took the paper test having better teachers. We find that the average 2014 percentile rank of teachers who would teach

students who took the PARCC paper test in 2015 was about 0-3 points higher in elementary school and 3-5 points higher in middle school relative to teachers whose students took the test online, suggesting the possibility of differential sorting by test mode. In the paper, we standardize scores within test mode so that, for example, students taking the paper PARCC test in 2015 have mean zero and standard deviation one. This effectively forces average value added to be equal across test modes. Results are similar when not standardizing across test modes with the exception of reading in elementary school, where the estimate of forecast deviation in Table 6 decreases from 35% to 20% when restricting the sample to PARCC paper test takers only.

New York City

We exploit the new statewide tests in grades 3-8 first implemented in spring 2006, which were accompanied by new standards in mathematics (there were no change in the ELA standards). Before 2006, the state tested only in Grades 4 and 8, but the district administered tests in Grades 3, 5, 6, and 7. In mathematics, we have value-added estimates for 2000 through 2010 for all grades. In reading, we have estimates from 2004 to 2010 for all grades except for fifth grade, where we also have estimates for 2003. During this time period, there were several changes in test administration. In 2003, the state shifted from item response theory to number-correct scoring. In 2010, the tests were moved from January to April. Finally, between 2003 and 2006, many English learners took an alternate test and were excluded from the main testing population. Our results are generally consistent when excluding years before 2003 or after 2010 or when excluding English learners.

North Carolina

North Carolina technical documentation made repeated note of the possibility of disruption in measurement caused by changes to tests and curriculum:

- "Test items will appear at that time to be more difficult than they will be when used operationally after the new curriculum has been implemented […] this kind of experience may follow any drastic change in the curriculum in any subject-matter area." (Math, Edition 2)

- "It is simply not possible to administer different tests, based on different curricula, in two successive years and expect the results to be in all senses as-expected." (Reading,

  Edition 2)

North Carolina constructs a developmental scale to measure growth from year to year in knowledge and skills. To determine a baseline for typical growth throughout the course of a school year, identical items are administered in adjacent grades (e.g., both third- and fourth-grade students are administered a set of items that would appear on the third-grade assessment). Scores are then standardized around fifth grade. For example, during Edition 1, fifth-grade reading and math scores ranged from 100 to 200 with mean 150 and standard deviation 10, by construction.

Data come from the North Carolina Department of Public Instruction, managed by Duke University's North Carolina Education Research Data Center. The data include student achievement on standardized tests in Grades 4 and 5 in math and reading from spring of the 1996-1997 school year through spring 2012.

Before 2007, North Carolina did not link students to teachers, but instead listed the proctor of a student's assessment. As in Xu et al. (2012), we attempt to restrict the sample to classrooms where the proctor is the classroom instructor by retaining a sample of classrooms where the characteristics of the test classrooms are similar to those in the instructional classrooms. We measure the mean squared difference between the instructional and test classrooms along percent male, percent White, and class size and keep self-contained classrooms with sufficiently small difference. In addition, we restrict our sample to classrooms with between 10 and 40 students and a majority of nonspecial-education students.

Washington

We obtain Washington student records from student longitudinal databases maintained by the Office of the Superintendent of Public Instruction. The state has required standardized testing in math and reading in Grades 3-8 since 2005-2006. For school years 2006 to 2009, the student data system included information on students' registration and program participation, but did not explicitly link students to their teachers. We therefore matched

these students to teachers using the proctor identified on the end-of-year assessment. The proctor variable was not intended to be a link between students and their classroom teachers, so this link may not accurately identify those classroom teachers. To ensure that these are likely to represent students' actual teachers, we limit the 2006-2009 sample to classrooms with between 10 and 33 students where the identified teacher is listed in the S-275 as 0.5 FTE in that school, teaches students in no more than one grade, and is endorsed to teach elementary education.[40] Since 2010, Washington data has included fields designed to link students to their individual teachers, based on reported schedules. However, limitations of reporting standards and practices across the state may result in ambiguities or inaccuracies around these links. That said, we identify math and reading courses using a combination of the Course Content Area code and string searches within the course names, and to guard against the possibility that elementary schools assign students to a "homeroom" teacher who does not actually provide math or reading instruction, we exclude teachers who teach "homeroom" courses at the elementary level. In our value-added models, we only consider students who are matched to exactly one math teacher and exactly one reading teacher using our matching system.

[40] Some of the data related to students and teachers used in this study are linked using the statewide assessment's "teacher of record assignment", a.k.a. assessment proctor, for each student to derive the student's "teacher". The assessment proctor is not intended to and does not necessarily identify the sole teacher or the teacher of all subject areas for a student. The "proctor name" might be another classroom teacher, teacher specialist, or administrator. For the 2009-2010 school year, we are able to check the accuracy of these proctor matches using the state's new Comprehensive Education Data and Research System (CEDARS) that matches students to teachers through a unique course ID. Using the restrictions described above, our proctor match agrees with the student's teacher in the CEDARS system for about 95% of students in both math and reading.

Appendix B – Accounting for Sampling Error in Adjacent-year Correlations

The results in this paper do not adjust for sampling error. When performing the adjustment described in this section, the basic patterns remain the same; importantly, the transitions with large drops unadjusted correlations continue to have large drops after performing the adjustment. Results are available from the authors.

To remove the teacher effectiveness measure not driven by random error, we adopt the correction described in Aaronson et al. (2007) and Goldhaber and Hansen (2013), who show that if estimated teacher performance consists of true performance and a random error term, then the correlation coefficient between the estimated performance of teacher *j* in two consecutive years can be written as the following:

$$corr\left(\hat{\tau}_{j,t}, \hat{\tau}_{j,t-1}\right) = \frac{cov\left(\tau^0_{j,t}, \tau^0_{j,t-1}\right)}{\sqrt{var\left(\tau^0_{j,t}\right) + var\left(\varphi_{j,t}\right)}\sqrt{var\left(\tau^0_{j,t-1}\right) + var\left(\varphi_{i,t-1}\right)}}.$$

In the above equation, $\tau^0_{j,t}$ represents true teacher performance and the denominator contains noisy measurements from both time periods. By removing the error variance, we estimate calculate the correlation of true performance over time:

$$corr\left(\tau^0_{j,t}, \tau^0_{j,t-1}\right) = \frac{cov\left(\tau^0_{j,t}, \tau^0_{j,t-1}\right)}{\sqrt{var\left(\tau^0_{j,t}\right)}\sqrt{var\left(\tau^0_{j,t-1}\right)}}.$$

To estimate $var\left(\varphi_{j,t}\right)$, we average the standard errors of teacher effects across all teachers. We then remove these random errors to calculate adjusted correlations.