The Common Core Conundrum: To What Extent Should We Worry That Changes to Assessments Will Affect Test-Based Measures of Teacher Performance?[1]

Ben Backes[2]
bbackes@air.org
James Cowan
jcowan@air.org
Dan Goldhaber
dgoldhaber@air.org
Cory Koedel
koedelc@missouri.edu
Luke C. Miller
lcm7t@virginia.edu
Zeyu Xu
zxu@air.org

October 2017

Abstract: Policies that require the use of information about student achievement to evaluate teacher performance are becoming increasingly common across the United States, but there is some question as to how or whether to use student test-based teacher evaluations when student assessments change. We bring empirical evidence to bear on this issue. Specifically, we examine how estimates of teacher value-added are influenced by assessment changes across 12 test transitions in two subjects and five states. In all of the math transitions we study, value-added measures from test change years and stable regime years are broadly similar in terms of their statistical properties and informational content. This is also true for *some* of the reading transitions; we do find, however, some cases in which an assessment change in reading meaningfully alters value-added measures. Our study directly informs contemporary policy debates about how to evaluate teachers when new assessments are introduced and provides a general analytic framework for examining employee evaluation policies in the face of changing evaluation metrics.

Citation: Backes, B., Cowan, J., Goldhaber, D., Koedel, C., Miller, L. C., & Xu, Z. (2018). The common core conundrum: To what extent should we worry that changes to assessments will affect test-based measures of teacher performance?. *Economics of Education Review*, *62*, 48-65.

[1] Acknowledgments: This work was supported in part by grant R305C120008 from the Institute of Educational Sciences to CALDER. We are grateful to feedback from Sean Corcoran, Tom Dee, Lars Lefgren, and discussants at the APPAM, AEFP, and CALDER conferences. We also thank our state partners for providing data access. Melanie Rucinski provided excellent research assistance.
[2] Corresponding author.

I. Introduction

Ongoing improvements in the capacity to store and analyze data have led to increases in the use of data-driven, outcomes-based metrics to evaluate the quality of services provided and worker performance in many professions. Examples of particular interest to the public include law enforcement, medicine, and education. The education sector has arguably been at the forefront among public employers in terms of using data to measure outcomes-based employee performance.

A focal outcome measure used in K–12 public schools is student achievement on standardized tests. Due in part to the increased availability of data systems developed in most states under the federal No Child Left Behind Act (NCLB), there is now a substantial body of evidence on the statistical properties of outcome-based measures of teacher performance, often referred to as teacher "value-added." Research has focused on issues such as the degree to which teachers differ from one another in their contributions to student achievement, whether value-added measures are biased, and the stability of the measures across time, test type, and model specification. Although there is ongoing scholarly debate about specific properties of value-added and how value-added measures should be used (e.g., see Corcoran & Goldhaber, 2013), there is consistent evidence that value-added is an informative measure of teacher quality. For example, several recent studies show that value-added is a strong predictor of future student outcomes by leveraging experimental and quasi-experimental variation in student-teacher assignments (Bacher-Hicks, Kane, & Staiger, 2014; Chetty, Friedman, & Rockoff, 2014a; Kane et al., 2013). Chetty, Friedman, and Rockoff (2014b) further link value-added to consequential longer-term outcomes such as wages, college attendance, and teenage childbearing. Other measures of teacher quality commonly used in teacher evaluations exhibit much weaker

relationships with student outcomes (Kane et al., 2011, 2013) and appear biased by teaching circumstance (Steinberg & Garrett, 2016; Whitehurst, Chingos & Lindquist, 2014). Teacher evaluations that incorporate value-added have a variety of potential policy applications, such as improving the targeting of retention/removal policies, compensation rewards, and professional development interventions.

Teacher performance evaluations that incorporate value-added have spread rapidly in recent years.[3] In New York City during the 2014–15 school year, for example, student performance on state test scores was formally incorporated into teacher evaluations and accounted for 20% of the total rating (classroom observations and other learning metrics accounted for the other 80%). A result is that the city's ratings became more evenly distributed relative to ratings in the rest of the state.[4] However, the use of value-added is controversial with teachers' unions and other groups opposing the incorporation of student-achievement measures into teacher evaluations (Darling-Hammond et al., 2012). Moreover, organizations such as the American Statistical Association (ASA) and American Educational Research Association (AERA), while not going so far as to oppose the measures, have urged caution in their use (AERA 2015; ASA, 2014).

The widespread implementation of the Common Core State Standards (CCSS), which entails changes to both states' educational standards and the associated student tests, has added

[3] We use the term "value-added" here as shorthand for measures of teacher performance based on student tests. Although the specifics of how the measures are calculated vary across states, they share common features (Goldhaber et al., 2014). Thirty-nine states and the District of Columbia now mandate that teacher evaluations include student growth measures (see Database on State Teacher and Principal Evaluation Policies, American Institutes for Research, Retrieved from http://resource.tqsource.org/stateevaldb/Compare50States.aspx; also see Steinberg & Donaldson, 2016).

[4] Disare, M., & Darville, S. (December 14, 2015). "92% of city teachers earn high marks in newest round of evaluations." *Chalkbeat New York.*

to the controversy.[5] A central objection to using test-based measures to evaluate teachers with the rollout of the CCSS is that it is unfair to hold teachers accountable for test results when new standards and assessments have been recently adopted.[6] 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 evaluated.[7] A related concern is that the curricular and testing transitions did not always occur simultaneously, creating potential misalignment between the curriculum and assessment.[8] Further complicating matters is that among states that originally adopted CCSS, to date, 10 have chosen to further revise their standards, which entails another round of rolling out new standards and associated assessments.[9]

In response to these concerns, in 2014, then Secretary of Education Arne Duncan granted a 1-year moratorium on the use of test-based metrics in teacher evaluations to states that had been required to incorporate them under their NCLB waivers.[10] A number of states delayed the incorporation of test-based measures of teacher performance into evaluations with the explicit reasoning that teachers need more time to prepare for shifts in standards and assessments, including but not limited to Colorado (Simpson & Torres, 2014), Pennsylvania (Chute & Niederberger, 2015) and Washington, DC (Brown, 2014). This did not mean that teacher

[5] Throughout the paper, we use the meaning of "standards" that refers to what students are expected to know rather than in reference to cut points on assessments.
[6] For example, see Chang, K.. (September 3, 2013). With Common Core, fewer topics but covered more rigorously. *The New York Times*, D2.
[7] For instance, AFT president Randi Weingarten argued that "the tests are evaluating skills and content these students haven't yet been taught." Source: Rose, M. (2013). "AFT calls for moratorium on Common Core consequences." *AFT News*.
[8] Polikoff and Porter (2014) study how the alignment of teacher instruction with standards and assessment content relates to teacher value-added and find a weak link (also see D'Agostino et al., 2007).
[9] Sawchuk, S. (2017). New York has rewritten the Common Core. Here's what you need to know. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/curriculum/2017/09/NY_replaces_common_core_here_are_the_details.html
[10] Announcement: http://www.ed.gov/blog/2014/08/a-back-to-school-conversation-with-teachers-and-school-leaders/. Note that the above-described issues were key concerns with the transition, but not the only concerns. A notable example unrelated to the curricular substance of the transition is that several states faced technical challenges with the rollout of computer-based Common Core tests (e.g., see Brown, 2016).

evaluations were not conducted at all, but rather that the weight on value-added was set to zero and hence the weights on other performance measures, such as observations of classroom practice, were increased.

Although the question of whether to use value-added during an assessment shift gained prominence due to the CCSS, changes to state educational standards and assessment regimes are quite common. For instance, the five states studied in this paper experienced 12 assessment changes in math and reading from 2000 to 2014, most of which have been accompanied by changes in standards. Indeed, in some states, the revision of standards and assessments is routine.[11] A notable difference between the CCSS and past changes is that in the CCSS era, many state and local education agencies are using, or are considering using, test-based measures of teacher performance as part of the formal evaluation process. Given the historical prevalence of changes to state standards and assessments and the increasingly common use of test-based measures of teacher performance in evaluations, the policy question of how to evaluate teachers during test regime shifts is likely to be salient for years to come.

Although it is not possible to know a priori the extent to which any specific test change will result in meaningful impacts on judgments about teacher performance, the fact that assessment changes are not new affords the opportunity to assess how past changes have affected value-added measures of teacher effectiveness. However, to our knowledge, there is no empirical evidence addressing this issue. We fill this gap in the literature, reporting on research assessing the extent to which value-added measures of teacher performance are affected by test changes. Specifically, we use longitudinal data from Kentucky, Massachusetts, New York City, North Carolina, and Washington state, each of which previously revised its student assessments, to

[11] North Carolina, one of the sites for this study, revised its standards and associated assessments on a recurring 5-year schedule, with a previous revision described as a "drastic change in the curriculum" (Bazemore et al., 2006).

explore the reliability and stability of teacher value-added during changes in assessment regimes.[12] The assessment changes at the sites we study occurred within the context of a wide variety of assessment and evaluation policies. We study two states that began assessing the CCSS before the introduction of the tests offered by the CCSS consortia (i.e., the Partnership for Assessment of Readiness for College and Careers and the Smarter Balanced Assessment consortia), three states that adopted new or revised learning standards that predate the CCSS, and two states that revised their assessments without altering the underlying learning standards.[13] The variation in these policy changes reflects the diversity of state experiences with respect to standards and assessment changes.

We begin our analysis by examining whether value-added estimates from assessment change years are more volatile because of the introduction of a new regime. In all of the math transitions we study, we find that value-added measures during assessment change years are similarly stable to measures from nontransition years. In reading our results are far more mixed, and at one site in particular—Kentucky—we observe a significant drop in the classification consistency of value-added corresponding to a test regime change. We also examine whether changes in teachers' rankings during assessment changes are associated with the characteristics of the students to which they are assigned. There is no evidence that volatility of teacher value-added during assessment changes is associated with student characteristics, nor is there any evidence that the rankings of teachers in disadvantaged classrooms are influenced by an assessment change.

[12] Our analysis is along the lines of what is advocated by McCaffrey (2013).
[13] In one state, Massachusetts, after 2 years of using state tests of Common Core standards, districts were recently given the choice (as of 2015) of whether to adopt PARCC or continue to use the state's existing CCSS-aligned test. This will present an interesting opportunity to study the transition to the PARCC test as more data become available.

We also apply the methods of Chetty et al. (2014a) and Bacher-Hicks, Kane, and Staiger (2014) to improve our understanding of the informational content of value-added during assessment changes. Specifically, we employ their framework to examine the extent to which value-added estimates from stable assessment years predict student achievement during assessment changes. Consistent with our finding of limited additional volatility of value-added during most assessment changes, we show that student achievement in regime shift years can generally be forecasted accurately, although not perfectly, by teacher value-added from stable standards and assessment regimes. The most notable exception is again in reading in Kentucky, where our ability to forecast student achievement during the assessment change using data from stable years is substantially degraded.

Overall, our findings indicate that although there is some degree of information loss in value-added estimates during assessment change years, much of the informational content of value-added persists. Following on prior studies showing that value-added is an informative measure of teacher productivity (Bacher-Hicks, Kane, & Staiger, 2014; Chetty, Friedman, & Rockoff, 2014a, 2014b; Kane et al., 2013), this implies that value-added from assessment change years is also an informative measure.[14]

Although the informational content of value-added is largely maintained during the test changes we study, we are unable to identify factors that can be used to predict instances of low informational persistence ex ante. This uncertainty can be viewed as a cost of using value-added during assessment change years and could contribute to calls for future moratoria. Within a broader cost-benefit framework, other factors that will contribute to the decision of whether to

[14] This summary of findings applies to value-added in math more than in reading; our mixed results in reading are consistent with other recent research showing that reading value-added is a less robust measure (Goldhaber & Hansen, 2013; Lefgren & Sims, 2012).

impose a moratorium in any particular locale include how value-added is used (e.g., how heavily is it weighted in teacher evaluations and how consequential are the evaluations in determining teacher outcomes ranging from employment, awards, tasks, and professional development) and the quality of available alternatives. A notable cost of a moratorium is that to the best of our knowledge, no other component of teacher evaluation systems in practice has been shown to be nearly as predictive of student success as value-added. Classroom observation scores are the most likely substitute in current systems and as noted above, in addition to being more weakly predictive of student outcomes, they seem biased in favor of teachers in low-socioeconomic status (SES) classrooms. Our finding that changes in the informational persistence of value-added during transitions are typically small, and moreover, even when persistence is meaningfully reduced value-added is still an informative measure of teacher quality, suggests that value-added continues to be an informative measure of teacher quality during standards and assessment transitions.

II. Background

A. Teacher Value-Added

There is a large literature on the analytic roots, statistical properties, and predictive validity of teacher value-added (see review by Koedel et al., 2015). In brief, value-added is a statistical measure meant to capture a teacher's unique contribution to student learning over the course of a year. Value-added models can be derived directly from structural cumulative achievement models under some assumptions, but the structural assumptions need not be satisfied in order for value-added to classify teachers accurately (Sass, Semykina, & Harris, 2014). A key challenge in estimating value-added is that students and teachers are nonrandomly sorted. The extent to which value-added models can account for the sorting is an empirical

question that has been studied extensively and recent, well-identified studies provide little indication that value-added measures are meaningfully biased (Koedel et al., 2015). For example, in a recent experiment conducted as part of the Measures of Effective Teaching Project (Kane et al., 2013), value-added estimated for teachers from observational data is shown to forecast student achievement in math accurately under random assignment of teachers to classrooms (the evidence for reading value-added is less compelling than for math but also consistent with forecast unbiasedness).

Chetty et al. (2014a) provide related quasi-experimental evidence on the informational value of value-added. They leverage arguably exogenous changes in teacher value-added at the school-by-grade level brought on by teacher mobility (also see Bacher-Hicks, Kane, & Staiger, 2014) as well as detailed data from the Internal Revenue Service (IRS) that is not commonly available to researchers. In both of these tests, these authors find that value-added estimates that account for students' prior test scores are forecast unbiased. In summary, although there remains some debate and research will undoubtedly continue to inform our understanding of value-added as a measure of teacher performance (e.g., see Chetty, Friedman, & Rockoff, 2017; Goldhaber & Chaplin, 2015; Rothstein, 2017), the strongest scientific evidence to date indicates that value-added is an informative measure. We take the evidence base on value-added as a point of departure for our study to examine how value-added in stable regimes differs from value-added in years when there is an assessment (and in some cases standards) change.

It is well-understood that value-added measures consist of persistent and nonpersistent components (e.g., Goldhaber & Hansen, 2013; Kane & Staiger, 2008; McCaffrey et al., 2009). The nonpersistent components reflect true year-to-year variability in teacher performance and estimation error. The persistent component is substantial and reflects what researchers typically

refer to as "teacher quality;" we use this same nomenclature. Teacher quality has been shown to be stable across settings in previous research. For instance, Chetty et al. (2014a) and Bacher-Hicks et al. (2014) find that value-added estimates for teachers who switch schools and grades exhibit no forecast bias in two different large school districts. Xu, Ozek, and Corritore (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 highly effective teachers continue to have positive effects on student achievement in math and reading when they transfer from low- to high-poverty schools.

In the case of an assessment change, at least three factors will determine how true teacher performance, which value-added aims to measure, will be affected. The first is the overall degree of change associated with the shift to new standards and assessments. This could result in changes to measured teacher performance as a result of shifts in the emphases of student mastery of some tasks over others on assessments, shifts that reflect the conscious choices of policy makers.[15] The more substantive the shift, the greater the potential for effects on measured teacher performance. Although anecdotally changes in standards and assessments are often viewed as being appreciable (e.g., see Bazemore et al., 2006) and can generate consternation in the education community, the degree of change is difficult to measure and we are not aware of any empirical work that aims to quantify the substantive importance of regime shifts. An advantage

[15] There could also be unexpected shifts in how students perform due to, for instance, the use of new test items used to access achievement.

of our study is that by pooling data from multiple sites and subjects simultaneously, we can examine a broad swath of changes as they occurred in practice.[16]

The second factor is teacher adaptability. When a change occurs, it could be that effective teachers are better able to adapt to the material emphasized on the new tests, especially when an assessment change is accompanied by a change in standards. This could make stability greater in transition periods. Alternatively, it may be that effective teachers are identified as such because they have built up specialized knowledge under a given test regime that cannot quickly be adapted to new tests or standards, and thus will be harmed by a switch. In this latter case, teachers with more experience under an old regime would be at risk of losing more test-specific skills when the test changes (see Chin, 2016).

A third factor relates to the extent to which student tests play to particular teaching strengths (e.g., instruction in long division). Research suggests that teachers are differentially effective at teaching different dimensions of a subject (Lockwood et al., 2007; Papay, 2011) and have different measured performance across tests (Corcoran et al., 2011; Papay, 2011). Thus, assessment changes that emphasize particular tasks could favor some teachers over others and lead to ranking changes among the workforce. To the extent that the new emphasis of some tasks over others reflects conscious choices by policy makers, these ranking changes may be desirable, but even then it is of interest to understand the disruption effect of a regime change.

In addition to factors that could influence true teacher performance, the statistical properties of value-added could also change with an assessment change. One possibility is that

[16] Associated with whether a change is "big" or "small" is the idea that some aspects of the change may happen slowly—for example, there might be a slow phase-out of old curriculum materials. In this and other similar examples, it is unclear how value-added measures will be affected. Using this specific example, if old curriculum materials phase out slowly, this could lead to increased stability of value-added during a transition because the change in conditions is smaller, or a decline in stability if the combination of new standards and old materials increases the importance of what we call "teacher quality" in determining student outcomes (see the next point on "teacher adaptability").

student performance on the old test may not be as predictive of student performance on the new test. This would weaken the predictive ability of the models used to estimate teacher value-added during the transition, and hence result in less precise estimates of teacher performance. This is potentially important as a reduction in precision implies an increased likelihood, for instance, that teacher classification into performance categories, based on value-added across different testing regimes, would be a reflection of statistical noise rather than true signal about teacher performance. Another possibility is that the targeting of the test changes, which would lead to changes to how well achievement is measured for different types of students, could in turn also influence estimates of teacher value-added (Koedel, Leatherman, & Parsons, 2012; Lockwood & McCaffrey, 2014; Stacy, Guarino, Reckase, & Woolridge, 2013).

Ultimately, the extent and direction of the effects of any changes in true teacher performance, and/or the statistical properties of value-added, coinciding with assessment changes are unclear a priori, which motivates our empirical investigation.

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 it is quite common for states to revise their standards and assessments. In this section, we briefly describe the standard and assessments changes we evaluate (see Appendix A for more details). Except for the recent changes involving CCSS, we are not aware of any research quantifying the substantive magnitude of each change, but by examining all of the changes using common methods, we capture the diversity of standards and assessment changes implemented in practice. Table 1 provides a summary of the changes covered by our study.

Kentucky: Kentucky adopted new, CCSS-aligned standards in 2010, becoming the first state to do so. During the 2010–11 school year, district leadership teams constructed student learning targets from the standards, which were then shared with teachers. The new standards were taught for the first time in the 2011–12 school year and students took new assessments for the first time in spring 2012. Schmidt and Houang (2012) identify Kentucky as one of the states with the highest degree of divergence between state standards and CCSS, suggesting that the adoption of CCSS was a meaningful change in standards.

The assessment in Kentucky before the adoption of the new test in 2012 saw a maximum score attained by many students (Innes, 2009; Koretz et al., 2014).[17] As shown in Section IV, estimates of teacher effectiveness in Kentucky are among the most volatile during the transition, and the properties of the prior exam could be a contributing factor. The skewness for the old Kentucky test, however, is only -0.35, which, based on results from Koedel and Betts (2010), should not be enough to cause substantial bias.[18]

A second issue with the Kentucky data is that the number of students identified as FRPL-eligible increased significantly and sharply, from about 12% in years prior to 2012 to about 60% in 2012 and following. The large, sharp increase in the FRPL share raises concerns about the reliability of the variable. Correspondingly, we do not control for FRPL in our models that estimate value-added for Kentucky teachers.[19]

---

[17] More than 12% of students earned the maximum score on Kentucky's older test; in contrast, on a typical exam in the remainder of the states in our sample, fewer than 3% of students earned the maximum score.

[18] As described in Appendix A, we implemented an alternative specification by probit transforming all pretest and posttest measures in Kentucky to examine this issue in more detail. However, this transformation makes very little difference in the results, with the exception of the estimate of forecast bias in the reading assessment (Table 6, column 5), which shrinks from 41% to 28% when performing the transformation.

[19] Our findings are generally similar in KY if we include the FRPL control, which is consistent with research showing that the most important controls in value-added models are prior test scores (Koedel et al., 2015). In cases where there is some discrepancy in the results depending on whether we include the FRPL control (most notably in Table 5 below, where the shift in the measurement of FRPL coincides with a transition), we prefer the models without the FRPL control because of the measurement issue.

Massachusetts: Massachusetts formally adopted learning standards that incorporated the CCSS in math and reading in 2010. The prior standards were first assessed in full in 2006 (Massachusetts Department of Elementary and Secondary Education, 2004a, 2004b). Massachusetts ranked 24th out of 50 states in Schmidt and Houang's (2012) measure of congruence between the prior state standards in math and CCSS, putting it in the middle among states in terms of the magnitude of the standards shift. Following the adoption of the CCSS in 2010, Massachusetts used an assessment focusing on areas common to the new and old sets of standards through 2012 (i.e., CCSS and the old standards). In 2013, Massachusetts began assessing the new standards using the state test in math and English language arts (ELA; for convenience, we use the terms "English and language arts" and "reading" interchangeably).

New York City: Before 2006, New York City (NYC) used its own tests for Grades 3–-8. In 2006, statewide testing was introduced in response to No Child Left Behind. The statewide tests replaced the NYC district tests and were accompanied by a standards change.[20] The state shared a toolkit with districts to support them in aligning their curriculum to the new standards.[21]

North Carolina: Before the adoption of the CCSS, curriculum revisions in North Carolina operated on a 5-year schedule and were planned well in advance. For example, the new K–12 math curriculum chosen in May 1998 was first used during the 1999–2000 school year. For the corresponding assessment, field test items were included as part of end-of-grade tests in the spring of 2000 and the new tests were introduced formally in 2001.[22]

[20] Introduction to the Grades 3–8 Testing Program in English Language Arts and Mathematics. http://www.scotiaglenvilleschools.org/parentinformation/38intro.pdf
[21] Kline, M. New York State Education Department forum on NYS learning standard for mathematics.
[22] The 1998 mathematics standard course of study and North Carolina mathematics tests. (2000). Public Schools of North Carolina.

Washington: The state of Washington introduced annual statewide testing in Grades 3–8 in spring 2006. The 2006 state assessments reflected a set of learning standards introduced in 2004 (Office of the Superintendent of Public Instruction, 2004a, 2004b). In July 2008, Washington released new *math*, but not reading, standards (Office of the Superintendent of Public Instruction, 2008), which were formally assessed for the first time during the 2009–10 school year. Also in 2009–10, the format of the state assessment changed in both math and reading.

Table 1 lists the assessment changes described in this section. The reading changes in New York City and Washington are the only cases where an assessment change was not accompanied by a standards change. For states where data from 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 student background characteristics, they also share commonalities in terms of their general content and structure. For instance, each state provides basic demographic and socioeconomic data for students, links between students and teachers, and identifiers that permit us to track teachers over time. Although nonrandom attrition of teachers in response to assessment or standards changes could potentially affect results, in an analysis omitted for brevity, we do not find any evidence

that teachers were more likely to exit in assessment change years, making it unlikely that compositional changes to the workforce around transitions influenced our findings.

Table 2 displays information for each state about the years and grades covered by our study, demographic information included, and numbers of students and teachers (detailed information about the construction of the analytic sample in each state can be found in Appendix A). As shown in the table, we focus on teachers in Grades 4–8 (4–5 only at some sites) who are responsible for math and reading instruction, which is typical of value-added research. Our results speak most directly to these teachers. In some states and school districts, growth measures are computed for teachers in other grades and subjects as well. Although our analysis does not directly examine teachers in these other grades and subjects, to the extent that our findings about the informational persistence of value-added reflect underlying teacher quality carrying over across tests, it would be reasonable to expect similar results.[23]

B. Estimating Teacher Value-Added

We perform the first portion of our analysis using estimates from a standard one-step value-added model (Koedel et al., 2015):

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

where A_{ijt} denotes achievement of student i taught by teacher j in year t, A_{it-1} prior achievement, X_{it} student-level demographic controls, and T_{jt} a vector of teacher-by-year indicator variables. We perform separate regressions for each subject, state, and year. The coefficients on the teacher indicator variables, β_{jt}, are estimates of teacher value-added. This

[23] At least when high-quality, centrally developed assessments are used. This statement is less likely to apply to, for example, teacher-generated assessments used for student learning objectives in some states (Lacireno-Paquet, Morgan, & Mello, 2014), about which we know relatively little. Teacher-generated assessments will be less responsive to changes in centrally driven curricular and assessment decisions (but more variable along many other dimensions).

model formulation is similar to value-added models used in teacher evaluations in several states and school districts. Estimates from this model are also highly correlated with estimates from other models, including models that do not include student-level demographic characteristics (Chetty, Friedman, & Rockoff, 2014a; Ehlert et al., 2014; Goldhaber et al., 2014).[24]

We control for prior test scores using a cubic polynomial in math and ELA as in Chetty, Friedman, and Rockoff (2014a). Chetty et al. (2014a) provide evidence that controlling for prior-year scores in this way removes nearly all of the bias from measures of teacher performance. Demographic controls for students vary somewhat among states (see Table 2 for a list of demographic information available in each state). We do not control for classroom characteristics in our models. Classroom controls cannot be separately identified from teacher-by-year effects in the absence of teachers who teach multiple classrooms, such as in the elementary samples we use for much of our analysis. Previous research suggests that excluding classroom-level controls will not affect our findings substantively.[25] Consistent with this expectation, when we add classroom controls to our tests for forecast bias below—which permit the inclusion of these controls—the results are very similar (results omitted for brevity).

We incorporate prior achievement as a control, rather than using a gainscore model (see Koedel et al., 2015 for more information on different value-added model specifications). The reason for this is that this specification of the value-added model has been demonstrated to

[24] We also briefly test whether estimates from a sparse value-added model without student covariates (i.e., controlling for same-subject prior year scores only) behave similarly to estimates from the VAM in Equation (1) during transitions. The sparse model alternative is of interest because some states use "student growth percentiles" to evaluate teachers, which do not condition on student characteristics. The results indicate that our findings of (a) comparable decile persistence rates in transition years relative to stable years (Table 3) and (b) teachers of disadvantaged students not being disproportionately harmed by transitions (Table 4) hold when using sparse VAMs in place of the richer model shown in Equation (1).

[25] 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. (2014) 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).

predict future performance well in experimental work (Kane & Staiger, 2008; Kane et al., 2013). Moreover, nonexperimental tests suggest that it estimates teacher effects with little bias (Chetty, Friedman, & Rockoff, 2014; Bacher-Hicks, Kane, & Staiger, 2014) and simulations find that it is more robust to nonrandom classroom assignment (Guarino et al., 2015b). As in other studies, we standardize test scores to have a mean of "0" and standard deviation of "1" within grade/subject/year.

Some investigations look into the properties of value-added use shrinkage estimators, where value-added estimates are shrunken toward the average to account for sampling error. We do not shrink our estimates. Guarino, Maxfield, Reckase, Thompson, and Wooldridge (2015a) show that shrinkage does not meaningfully change reliability and it makes little difference in the rank ordering of teachers. Consistent with their findings, none of our results are affected substantively by whether we explicitly account for estimation error in our value-added estimates. We show this in Appendix B using an ex post adjustment to account for estimation error.

Following on the discussion in Section II.A, there are several reasons teacher value-added may be sensitive to changes in testing regimes. To formalize ideas, consider a model of the components of estimated teacher effectiveness that builds on Goldhaber and Hansen (2013) and Winters and Cowen (2013):

$$\hat{\beta}_{jt} = q_j + \delta_{jt} + \tau_j^k + \varepsilon_{jt}. \quad (2)$$

In the above equation, $\hat{\beta}_{jt}$ is estimated value-added for teacher j in year t, q_j represents persistent teacher quality, δ_{jt} year-to-year shocks to performance (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 assessment changes, leaving the latter two terms as channels

through which estimated value-added can be affected by regime shifts.[26] Connecting equation (2) back to the discussion in Section II.A, the term τ_j^k captures differences in the ability of teachers to adapt to new tests and/or the influence of changes to test content that emphasize different dimensions of quality that may vary within teachers. The term ε_{jt} captures the influence of changes to the statistical properties of value-added deriving from changes to the level of the test, model fit, etc.

We cannot definitively distinguish between the potential reasons why value-added might change across test regimes so the results we present below are best viewed as encompassing changes in value-added measures that could be related to several factors. However, in Section V we argue that to the extent that measures of teacher quality change during an assessment shift, changes to test content are more important than changes to test measurement error and model performance.

C. Measuring the Stability of Teacher Performance During Regime Changes

Denote the vector of student test scores for students taught by teacher *j* in time *t* under assessment regime A as A_{jt}^A, and under regime B as A_{jt}^B. Let *AA* denote value-added calculated using year *t* and year *t-1* tests from regime A:

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

and let *BA* denote value-added calculated using a prior-year test from *A* and a post-test from *B*:

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

[26] We again note the aforementioned possibility that the signal of teacher quality becomes more prominent during a transition, which would manifest itself in the variance of q_j, and possibly δ_{jt}, rising relative to the variance of the other components (namely ε_{jt}). All else equal, if transitions increase the variance share of q_j, this would increase the year-to-year correlations in teacher value-added that involve an assessment-change year.

We call years when the pretest and post-test are from the same regime "single-test years" and years when the pretest and posttest are from different regimes "multitest years."

To illustrate how we examine stability when assessments change, take the following example that shows the regimes for the first assessment change in North Carolina (in math), with 2001 (spring) being the initial year of a new math assessment. Solid dots in Figure 1 indicate tests from regime *A* and empty dots indicate tests from regime *B*:

Figure 1: Example of Stable and Transition Years

	1998	1999	2000	2001	2002	2003	2004
Post-test	●	●	●	○	○	○	○
Pre-test	●	●	●	●	○	○	○
Type	*S*	*S*	*S*	*M*	*S*	*S*	*S*
		Stable		Transition		Stable	

Transition years are those where either the current or prior year is a multitest year, thus there are two transition years associated with each multitest year. We analyze the stability of teacher rankings by showing 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. Consider the correlation between estimated teacher value-added in 2000 and 2001. The estimates from 2000, a single-test year ("S"), are calculated using year *t* and *t-1* tests from the same regime (regime *A*), but 2001 is a multitest year ("M"). We describe the correlation between value-added in 2000 and 2001 as a *transitional measurement* because it involves a multitest year. In contrast, the correlation between value-added in 1999 and 2000 is a *stable measurement* because both years are single-test years. We document the extent to which the stability properties of value-added differ by whether the measurement period is transitional versus stable.

We formalize and extend our analysis by regressing the absolute value of the change in a teacher's percentile rank between years *t-1* and *t* on classroom characteristics and whether the year-to-year change is a transitional measurement:

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

In equation (5), $Char_{jt}$ contains teacher *j*'s average class values of %Black, %FRPL, 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 in a multitest year.[27] A positive value of α_1 would indicate that transitional measurements are associated with increased volatility of teacher rankings relative to measurements from stable standards and assessment regimes.

Equation (5) is useful to assess the overall change in the volatility of teacher rankings associated with a test transition, but it does not examine systematic effects for teacher subgroups. For example, it may be that volatility increases more for teachers in some types of classrooms than others. To examine this possibility, we expand the model 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 β_2 measures the extent to which assessment changes differentially affect teachers who differ by the characteristics of their students. 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. Forecasting Value-Added During Assessment Changes

We complement our analysis of the properties of teachers' annual value-added estimates as described thus far with tests for whether value-added exhibits the statistical property of

[27] An alternative approach would be to define a transition year to be solely the initial year after a regime change (i.e., in the above example for North Carolina, only the 2000-to-2001 correlation would be defined as "transitional"). Results are similar to what we report below if we restrict our attention to the initial year after a regime change.

"stationarity" through the assessment changes we study. The property of stationarity would imply that the informational content of value-added persists through an assessment change. The tests for stationarity are designed to fit our application but follow from the work of Chetty et al. (2014a). Specifically, we construct forecasts of student achievement during multitest years that rely on data from single-test years only, and formally build into the forecasts the assumption that value-added is a stationary process (the stationarity assumption is a key part of the Chetty, Friedman, and Rockoff procedure). Forecasting inaccuracy within this framework is indicative of stationarity being violated.

To implement the stationarity tests, we first construct value-added measures in single-test years using the forecast-based approach of Chetty et al. (2014a). These measures depend only on value-added estimated from other single-test years. We start by creating a residualized test score for each student *i* in year *t* using the following regression:

$$A_{it} = \alpha_j + \delta A_{it-1} + \gamma X_{it} + \varepsilon_{it}, \quad (7)$$

where A_{it} represents test scores for student *i* in year *t*, A_{it-1} a vector of prior year scores, X_{it} contains the same demographic information used in our earlier value-added models, and α_j is a vector of teacher fixed effects. With $\hat{\delta}$ and $\hat{\gamma}$ obtained from estimating equation (7) we produce residualized student scores:

$$A_{it}^{*} = A_{it} - \hat{\delta}\, A_{it-1} - \hat{\gamma}\, X_{it}. \quad (8)$$

A_{it}^{*} thus represents residual test scores after adjusting for student demographics and prior test scores. These equations follow Chetty et al. (2014a) and offer the conceptual benefit of

identifying the parameters in the δ and γ vectors using within-teacher variation (because δ and γ are estimated with teacher fixed effects included in equation 7).[28]

Teachers' value-added forecasts are constructed by first averaging the residual scores for the students assigned to each teacher to obtain mean residual test scores for each teacher *j* in year *t*, $\bar{A}_{jt}$. Forecasting coefficients, ψ_s, where *s* denotes the distance in time from the forecast target (period *t*), are then estimated to minimize the mean squared error of test-score forecasts as follows (see detailed description in Chetty, Friedman, & Rockoff, 2014a):

$$\psi = arg\ min_{\{\psi_s\}} \sum_j \left(\bar{A}_{jt} - \sum_{s \neq t} \psi_s \bar{A}_{js}\right)^2. \quad (9)$$

The estimates of ψ_s obtained from equation (9) are then used to forecast value-added for each teacher *j* in year *t* using data from teacher *j* in available years outside of *t*:

$$\hat{\mu}_{jt} = \sum_{s \neq t} \psi_s \bar{A}_{jt}. \quad (10)$$

The estimate $\hat{\mu}_{jt}$ is designed to minimize prediction error and is implicitly shrunken as noted in Chetty et al. (2014a). Thus, no additional ex post adjustments for estimation error are necessary.

The forecasting coefficients in the vector ψ_s are the key to our tests of stationarity. These coefficients are estimated entirely using data from stable regimes and indicate the predictive power of value-added from outside years. Our tests for stationarity through transitions ask how well we can predict value-added during a multitest year using the predictive coefficients ψ_s. If value-added is stationary during assessment shifts, which implies persistent informational content, the predictive power of outside-year value-added should not change during multitest years and the predictive coefficients ψ_s should be effective for forecasting purposes. The more that our forecasts deviate from actual, residualized student achievement during multitest years,

[28] Although as a practical matter, Chetty, Friedman, and Rockoff (2014a) show that this feature is not critical to the procedure.

the larger the implied change in the informational content of value-added as a measure of teacher effectiveness. Formally, we implement the tests using the following forecasting regression:

$$A_{it}^{*} = Y_t + \lambda \hat{\mu}_{jt} + \xi_{it}, \qquad (11)$$

where Y_t is a year-fixed effect. Under the stationarity assumption, the OLS regression in equation (11) should yield an estimate of λ that is indistinguishable from one because $\hat{\mu}_{jt}$ is the best linear predictor of A_{it}^{*}.

We illustrate with an example of a hypothetical regime shift, shown below, where the first 3 years use old assessments and the last 4 years use new assessments, with period 0 below indicating the initial year of the new assessment. This facilitates the estimation of 3 years of value-added from single-test years in the preperiod, 1 year of multitest value-added, and 3 years of value-added from single-test years in the postperiod.

Figure 2: Hypothetical Regime Shift with Three Stable Years Surrounding Multitest Year

	−3	−2	−1	0	1	2	3
Post-test	●	●	●	○	○	○	○
Pre-test	●	●	●	●	○	○	○
Value-added estimate	S	S	S	M	S	S	S

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 (Chetty, Friedman, & Rockoff, 2014a). So, for instance, the predictive validity of teacher value-added period 1 over value-added in period 3 will be the same as the predictive validity of value-added in period -3 over value-added in period -1. To test for a violation of stationarity during the multitest year, we use the single-test years to obtain the forecasting coefficients ψ_s from equation (9) for all values of $|t-k|$, where $k \neq t$. In our example, this would yield predictive coefficients for $|t-k|=1$, $|t-k|=2$, and $|t-$

$k|=3$ that can be used to generate the forecast for the multitest year (note that we can obtain predictive coefficients for larger values of $|t-k|$ in this example, but the largest value of $|t-k|$ that separates the multitest year from a single-test year is 3). Denoting these coefficients as ψ_1, ψ_2, and ψ_3, predicted teacher value-added during the multitest year, τ, is based on the forecasting coefficients as follows:

$$\hat{\mu}_{j\tau} = \hat{\gamma}_0 + \psi_1\overline{A}_{j\tau-1} + \psi_2\overline{A}_{j\tau-2} + \psi_3\overline{A}_{j\tau-3} + \psi_1\overline{A}_{j\tau+1} + \psi_2\overline{A}_{j\tau+2} + \psi_3\overline{A}_{j\tau+3} \qquad (12)$$

In equation (12), $\bar{A}_{jt}$ is defined as above as the average residualized achievement for students of teacher j in year t. The equation exemplifies how the value-added forecasts from equation (10) are applied to a multitest year. In cases with a different number of single-test years near the transition than in our example, equation (12) is modified to use all available information.

After obtaining the forecasts for each teacher j, $\hat{\mu}_{j\tau}$, we then estimate equation (11) using residualized student achievement from the multitest year only as the dependent variable. The predictor of interest in equation (11) for the multitest year, $\hat{\mu}_{j\tau}$, is interpreted as the best linear predictor of $A^*_{i\tau}$ under the assumption that stationarity of teacher value-added is upheld through the multitest year. An estimate of λ in equation (11) indistinguishable from one is consistent with value-added being a stationary process through the transition. Note that stationarity can be upheld even if value-added estimates are biased (e.g., persistent sorting bias could lead to stationary value-added estimates). Additional tests such as the ones performed in Chetty et al. (2014a) and Bacher-Hicks et al. (2014) are necessary to directly test for bias. We use evidence from Chetty et al. (2014a) and Bacher-Hicks et al. (2014), with both studies showing that the

scope for bias in similar value-added estimates is small, as a point of departure for our study of how the measures change during test transitions.[29]

Finally, we note that although the Chetty et al. (2014a) framework is useful for assessing the statistical property of stationarity and is complementary to the rest of our analysis in this way, we do not use the value-added forecasts developed by Chetty et al. (2014a) in other parts of our study for two reasons. First, forecast-based value-added measures along the lines of those used by Chetty et al. (2014a) are not used in any policy application of which we are aware. In contrast, annual value-added measures are used in several active policy applications including the District of Columbia and Pittsburgh, among others. Second, as will become clear below, the Chetty et al. (2014a) procedure estimates teacher value-added in any given year as a function of data from other available years. This makes it less useful in our examination of how the properties of value-added change from year to year because value-added measures for teachers in any given year pair are forecasted with overlapping information. Using the Chetty et al. (2014a) approach in the preceding analysis would lead to an overstatement of, for example, correlations in estimated value-added across adjacent years relative to what would be expected in the typical policy application that relies on single-year measures (Schochet & Chiang, 2013).

IV. Results

A. Correlation of Value-added Across Test Changes

Figures 3a–3g show adjacent-year value-added correlations. Vertical bars mark multitest years. In general, the correlations are in line with what has been shown in previous research (e.g.,

[29] The tests in Chetty, Friedman, and Rockoff (2014a) and Bacher-Hicks, Kane, and Staiger (2014) that address the scope for bias use a teacher-switching quasi-experiment that examines how changes in student test scores in school-by-grade cells align with changes in forecasted teacher value-added in the same cell. The teacher-switching test is not feasible to implement across the different testing conditions we study because it cuts the data too thinly, with the result that the test for bias is too imprecise to be informative (the way the test is structured, the null hypothesis is forecast-unbiasedness; imprecise tests that fail to reject the null are not useful). Additional information is available from the authors upon request.

Chetty, Friedman, & Rockoff, 2014; Goldhaber & Hansen, 2013; Kane & Staiger, 2011; McCaffrey et al., 2009), although there is some cross-state variation. The correlations are higher in math than in reading in most years and most states, but again, differences between the two subjects vary across states.[30] We focus our attention on whether the correlations change during transitions, rather than the implications of the overall level of time persistence in value-added.[31]

Conceptually, for multitest years, a reason to expect the first correlation—between value-added in the initial stable regime and the multitest year—to be lower is that the outcome changes in the multitest year (i.e., to the new test). For the second correlation—between value-added in the multitest year and value-added during the first year of the *next* stable regime—a reason to expect a lower correlation is that the control variables are from the old test during the multitest year. In practice, for the most part, the correlations that involve multitest years are very similar to correlations for single-test years. The exception is Kentucky, where the correlation between value-added in the year before the new test and value-added in the multitest year drops noticeably in math and even more so in reading. The correlation fully rebounds in math after the transition, but the posttransition correlation in reading remains lower. A possible explanation is that the change in assessments in Kentucky represented a more significant difference between the earlier and later student outcome tests than was the case in the other sites (Schmidt & Houang, 2012).

[30] 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 once the adjacent year correlations have been adjusted for measurement error (see Appendix B for details).

[31] Numerous studies show that value-added measures are sufficiently precise under normal circumstances to potentially improve the quality of the workforce by informing personnel decisions (Condie, Lefgren, & Sims, 2014; Goldhaber, Cowan, & Walch, 2013; Rothstein, 2015; Staiger & Rockoff, 2010; Winters & Cowen, 2013). The year-to-year persistence of value-added has also been shown to be similar to the persistence of performance measures for other professionals ranging from insurance salespeople to athletes (McCaffrey et al., 2009).

B. *The Tails of the Value-Added Distribution*

A primary objection to teacher evaluations that incorporate information from new assessments is that teachers are not prepared for the new tests and it is unfair to use them in personnel evaluations without allowing teachers time to adapt. Although the correlations presented above are informative about this issue, teacher evaluation systems implemented in practice have focused primarily on identifying teachers in the tails of the quality distribution for high-stakes intervention (e.g., see Dee & Wyckoff, 2015; NCTQ, 2016). With this policy context 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 rankings during stable and transition periods. Specifically, we take teachers whose value-added placed them in the top/bottom 10% in year *t-1* and measure the share who remain in the top/bottom 10% in year *t* for teachers who are observed in both *t-1* and *t* under each type of testing condition.

Our results are presented in Table 3. The decile-persistence levels based on single-year value-added measures are somewhat volatile even in stable years, a finding that is consistent with previous research (e.g., Schochet & Chiang, 2013). Also consistent with previous research (e.g., Goldhaber & Hansen, 2013), in most cases rankings in stable regimes tend to be more volatile at the bottom of the distribution than the top, reflected in Table 3 by a general pattern of smaller persistence in the lowest decile relative to the highest decile.[32]

Of interest for our investigation of assessment changes is that persistence in the tails is generally very similar regardless of whether there is a transition, which mirrors the correlations

[32] Table 3 is not structured to paint a thorough picture of the volatility. For example, the decile persistence measures do not distinguish between teachers who move from, say, the bottom-to-second decile from teachers who move from the bottom-to-top decile (i.e., we use a binary coding where both of these instances would be treated equally as "not persistent"). Other studies have looked in more detail at the issue of large versus small moves in teacher value-added rankings over time (e.g., see Aaronson, Barrow, and Sander, 2007).

in Figures 3a–3g. Each stable regime is associated with a likelihood 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. For example, for math teachers in Washington, 38% and 36% of teachers who were in the top decile in 1 year remained in the top decile the following year during the first and second stable regimes, respectively. During this transition, the share was 38%. The exception to this pattern of results is again Kentucky in reading, where the transition period is marked by excess instability.[33]

To elaborate further, in only 5 of 24 cases in Table 3 in which we observe a stable regime before and after a transition (i.e., the 24 refers to all sites except MA) is the classification consistency in transition periods more than 1 percentage point below the minimum of the values given by the two surrounding stable regimes. Four of the five cases are in reading (the five cases are: Kentucky top decile math, Kentucky bottom decile reading, NYC elementary top decile reading, and both the top and bottom deciles in Washington for reading). In only one case out of 24—the bottom decile for reading in Kentucky—is the difference between the transition value and the minimum of the surrounding stable values statistically significant. In some instances, we even observe higher classification consistency during transition periods (e.g., the top decile for North Carolina's first math transition), at least nominally, although this is rare.

Although this paper focuses on estimates of teacher performance in transition years, states also use school-level growth measures in accountability systems. A potential political middle ground for accountability during transition years may be to hold schools accountable for student learning, but not teachers. In results available from the authors, we find that school-level

[33] Recall from Section III.C that each assessment change corresponds to two transitional measurements. Results are similar when estimating the transition likelihoods of the first and second transition years separately. Results are available from authors.

value-added exhibits similar patterns as the teacher-level value-added results presented here in that school-level decile persistence rates in transitions are consistent with the surrounding stable regimes.

C. *Teacher Rankings By Classroom Type in Single- and Multitest Years*

As discussed above, a number of factors associated with any particular regime shift can potentially influence how teachers in different types of classrooms are affected. In Table 4, we examine how teachers are ranked based on value-added in single-test and multitest years for three types of classrooms using definitions from Goldhaber et al. (2014): advantaged classrooms, which fall into the top quintile of prior year achievement (averaged across math and reading) and the bottom quintile of percent FRPL for a given year; average classrooms, in the middle quintile of prior year achievement and percent FRPL; and disadvantaged classrooms, in the bottom quintile of prior year achievement and highest quintile of FRPL students.[34] Our approach is straightforward: we estimate teacher value-added for each year in each state for all teachers and then report the average percentile rank of teachers by classroom type from single-test and multitest years.

We find little evidence that teachers in disadvantaged classrooms are disproportionately harmed, on average, in multitest years. If anything, the average value-added percentile ranking of teachers in disadvantaged classrooms is somewhat higher in many of the multitest years we observe. Thus, we conclude that teachers placed in disadvantaged classrooms do not fare worse on value-added measures in multitest years relative to single-test years.

[34] As noted above, the FRPL measure in Kentucky varies wildly over time (for example, it exhibits a 1 year jump from 12% in 2011 to 61% in 2012) so we use only classroom achievement to define classroom advantage for that state. However, results are substantively similar when we do classify classrooms using FRPL.

D. *Regressions Predicting Change in Volatility of Teacher Rankings*

We formally test whether transition periods are associated with greater volatility in teachers' rankings by regressing the absolute value of the change in each teacher's percentile rank from year *t-1* to year *t* on classroom characteristics and whether either year is a multitest year (i.e., whether the ranking change is a transitional measurement). If switching to a new assessment is associated with increased volatility, one would expect a positive and significant coefficient on the transition indicator. Results are shown in Table 5. Odd-numbered columns show a base specification with no interaction terms between the transitional measurement indicator and other explanatory factors.

In terms of overall patterns, teachers in classrooms with high percentages of FRPL and Black students tend to have slightly more volatile rankings at most sites. In all instances, teacher rankings for reading 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 transitional-measurement indicator in all states but Kentucky and Massachusetts (in elementary only) represent changes of less than 1 percentile point, and are even negative in some cases. In states other than Kentucky, these results suggest that value-added is not substantially more volatile during transitions than in other years. However, Kentucky shows a moderate increase in volatility of about 10% relative to baseline (i.e., a change 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 sites other than Kentucky and Massachusetts elementary schools show an increase in volatility of more than

5%. This is well within the year-to-year variation we observe in stable standards and assessment regimes. The addition of the interaction terms in even-numbered columns provides no new insights aside from evidence that the volatility in teacher rankings during the transition in Kentucky is largely driven by reading.[35] Overall, the results from Table 5 indicate that test transitions are not typically associated with meaningful changes in the volatility of teacher rankings.

E. *Estimates of Forecast Instability*

Table 6 displays the results of the forecasting tests described by equation (11). The key feature of these tests is that we do not use any data from multitest years in constructing the forecasts. Before turning to our analysis of transitions, we begin in column (1) by showing results that are entirely internal to single-test years; that is, we use value-added from single-test years to forecast student achievement in other single-test years. If value-added is stationary during stable testing regimes at our sites, an expectation is that the forecasting coefficients in column (1) should be equal to 1.0. Consistent with this expectation, in three of the seven state-by-schooling-level analyses, we cannot reject the null hypothesis that the forecasting coefficient is 1.0 (Kentucky, Massachusetts elementary, North Carolina); and while in the other four cases our coefficients are statistically distinguishable from 1.0 (Massachusetts middle, New York City elementary and middle, Washington), the maximum deviation of the estimated forecasting coefficient from 1.0 is 0.044 (Massachusetts middle) and the smallest deviation is just 0.023 (New York City elementary). Our summary interpretation of these results is that value-added is a stationary or near-stationary process at the sites we study based on stable-regime data.

[35] In results available from the authors, we also find no evidence of increased volatility in transition years for teachers who were in the top or bottom quintile in the prior year value-added distribution.

Next, we turn to the question of whether test regime changes disrupt the stationarity of value-added. To answer this question, we use the forecasting coefficients to forecast value-added out of sample during multitest years. If the same level of stationarity observed during single-test years carries through multitest years, we would expect similar forecasting coefficients to what we show in column (1).

The results for all multitest years at each schooling level, pooled across both subjects, are shown in column 2. Although the pooled coefficients in column 2 mostly remain close to 1.0, we can statistically reject a value of unity in six of the seven cases and on average, the estimates in column 2 are further from 1.0 than their analogs in column 1. In columns 3–6, we disaggregate across subjects and assessment changes to gain further insight. The disaggregated results reveal that the largest forecasting deviations (i.e., the estimates that imply substantial breaks from stationarity during the multitest year) occur for reading. Most notably, the forecasting deviation in Kentucky for reading is particularly large—41.4 percent—and in Washington, the forecasting deviation in reading is almost 16% (15.9). In contrast, in math, only one forecasting deviation during a multitest year exceeds 10% (12.2%; New York City middle schools).[36] Interestingly, these represent very different kinds of states: Kentucky had both a change in standards and assessments so we might expect larger forecasting deviations, but Washington only had a change to its assessment (in the case of reading). Obviously, we cannot infer too much about how likely it is for forecasting deviations to be related to the type of changes states make from just two states, but the Washington findings show that even assessment changes alone can lead to forecast deviations.

[36] When plotting the relationship between test score residuals and forecasted value-added in multitest years in a binned scatterplot as in Chetty et al. (2014a), the estimated relationship appears to hold at all points in the test score distribution, suggesting that the accuracy of the forecast does not depend on whether a teacher is predicted to perform in the middle or tails of the distribution.

Overall, cases of forecast deviations in excess of 10–15% in Table 6 are infrequent. Although the table shows that teacher performance in multitest years cannot typically be forecast with the same accuracy as in single-test years, suggesting a modest disruption of stationarity, each of the multitest coefficients is very far from zero. This indicates that measured teacher performance in single-test years is highly predictive of, and thus strongly related to, measured performance in multitest years.[37]

Finally, another way to conceptualize the informational content of value-added in multitest years is to consider the prediction of teacher effectiveness in the year following a multitest year. In results omitted for brevity, we find that including the multitest year improves forecasting accuracy relative to excluding it, even during the reading transitions where our forecasts for the multitest year are least accurate. There are two reasons that including the multitest year leads to better forecasts of subsequent years. First, including the measure from the multitest year improves precision by incorporating additional data that the analysis thus far has shown to be informative. Second, using the multitest year improves the forecast by adding information from the new test, thus down-weighting the contribution of the old test to the forecast of posttransition value-added (although this second benefit is marginal for most transitions).

V. Discussion

The largest cases of volatility in teacher rankings and departures from forecast stability occur during reading transitions. One potential explanation for this pattern is that the typical scope for revising reading standards/tests is larger than for math. We find some evidence of this

[37] The results in columns 2–6 of Table 6 use data from single-test years before and after multitest years (when available) for the out-of-sample forecasts. Our findings for forecasts during multitest years are qualitatively similar if we use data from solely before the multitest year to construct the forecast.

when benchmarking revised assessments in Kentucky against a nationally administered test that did not change when Kentucky adopted the new state assessments. Figure 4 shows correlations between student scores on Grade 8 end-of-grade tests in Kentucky, which were changed in 2012, and the ACT-administered EXPLORE test, which was not, for both math and reading. As shown in the figures, while the math test has roughly the same correlation with EXPLORE through the transition, this is not the case in reading. We take this as suggestive evidence that the reading test in Kentucky changed more than in math.[38]

For math in Kentucky, we see the year-to-year correlation of value-added drop in the multitest year (Figure 3) without a corresponding drop in forecast accuracy (Table 6). Two factors help to explain this apparent puzzle. First, while clearly observable, the correlation drop during the multitest year in Kentucky for math is much smaller than the drop for reading: math value-added still carries a relatively persistent signal through the transition. Second, unlike in reading, the year-to-year correlation in math rebounds after the transition. The failure of the reading correlation to rebound violates the assumption of stationarity since the one-year-apart correlations are clearly not constant over time. A change in the year-to-year correlation of value-added after a transition is also visible to a lesser extent in Washington's reading transition (Figure 3g), for which we also document lower forecasting accuracy in Table 6.

With regard to the substantial increase in volatility of reading value-added during the Kentucky transition we consider, and subsequently rule out, three potential explanations. First, whether measured by density divergence between old and new tests (see Frölich, 2004) or skewness of old and new tests (Koedel & Betts, 2010), we find that the distributional properties

[38] Although Grade 8 students in Kentucky are not included in the preceding analyses as our main analysis sample contains only elementary school students, this exercise allows us to examine changes in the Kentucky end-of-grade test relative to a stable test.

of math and reading tests are generally similar in Kentucky, and thus distributional changes are not a probable explanation for why reading tends to have a higher degree of instability. Second, related to the possibility mentioned above that bad model fit during a multitest year might add additional noise and alter teacher rankings, we examine the predictive power of prior test scores over current test scores at the student level, but there is no change in the predictive power in reading during the transition that can account for our findings.[39] Finally, more generally, we rule out excessive estimation error during the multitest year in Kentucky as an explanation by showing that when we adjust the year-to-year correlations to remove the influence of estimation error (described in Appendix B), the patterns of results presented above remain.

VI. Policy Implications and Conclusions

Teachers' contributions to student performance on state assessments are being used, or under consideration for use, as a formal part of multipronged evaluation processes in many states and school districts across the United States. Indeed teaching is just one of several professions of particular interest to the public where outcomes-based, data-driven measures of efficacy are increasingly common.[40] The education sector is arguably at the forefront in developing and using these types of measures in policy applications. Our study aims to understand how changes to state standards and assessments, which affect the student-performance metrics used to evaluate teachers, affect teacher rankings.

An impetus for our study is the policy attention this issue has received as a result of the rollout of the CCSS (e.g., the CCSS rollout resulted in the aforementioned 1-year moratorium on

[39] In results available from the authors, we find that Spearman rank coefficients of correlations between student test scores in *t* and *t-1* are very similar in single-test years and multitest years.

[40] As noted above, other professions of interest include medicine and law enforcement. For example, Dimick et al. (2009) construct outcomes-based measures of hospital efficacy that are in the same spirit as value-added (given their focus on hospitals, the analogous level of analysis in education would be a school).

the use of test-based teacher evaluations granted by the United States Department of Education in 2014). Moreover, as illustrated by our study, state standards and assessment changes historically are quite common and we have every reason to expect more changes in the future. Although it is not possible to predict with certainty how shifts to CCSS-aligned standards and assessments, and other future shifts, will affect the quality of the information contained by estimates of teacher value-added, our investigation of prior standards and assessment changes provides insights to help guide policy decisions about teacher evaluations during transitions.

Our findings indicate that previous assessment changes have typically had minimal effects on the stability of estimated teacher value-added and teacher rankings in the tails of the distribution. In addition, in most cases we find that teacher performance in stable regimes forecasts student test scores during transitions well, although not as well as during nontransition years. Unlike in math, where we find that value-added measures from stable and transition periods are similar throughout, in reading, there are several cases where the informational content of value-added is substantially altered during a transition. 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 the underlying quality measures more generally (i.e., teacher value-added in reading has been shown to be a less informationally robust measure than in math in previous research; see Goldhaber & Hansen, 2013; Lefgren & Sims, 2012). However, despite the fact that during some transitions the informational content of reading value-added is clearly degraded, in no instance do the measures cease to be informative about teacher performance. Even in the most volatile case (Kentucky), an increase of one standard deviation in forecasted

student test scores—as forecasted by teacher performance in stable years—is associated with a 0.57 standard deviation increase in observed scores during the multitest year.

In summary, our findings point broadly toward value-added continuing to be an informative measure of teacher effectiveness during assessment changes, particularly in math. A moratorium on the use of value-added during transitions would discard information that is nearly as informative about teacher performance as similar measures from nontransition years in most cases. That said, our analysis suggests that it will be difficult ex ante to identify situations where the informational content of value-added changes meaningfully with a test regime change and this uncertainty could cause angst among the workforce, contributing to calls to halt test-based evaluations during transitions, especially when high stakes are attached. This policy decision must be weighed against the alternatives, which include either a moratorium on evaluations entirely or a shift in weight toward non-test-based measures. Previous moratoria suggest that the latter option is the more likely alternative; whether it is preferable depends on the relative quality of non-test-based measures (which are less informative and likely biased per Kane et al., 2011, 2013; Steinberg & Garrett, 2016; Whitehurst et al., 2014) and their own informational robustness to test-regime changes (we are not aware of any empirical evidence on the second issue).[41] We conclude by noting that suspending the use of test-based metrics in teacher evaluations during assessment changes is equivalent to treating their informational content as null, which from a purely analytic perspective is not supported by our analysis.

[41] Another rationale for a moratorium is that it can help achieve political objectives and may thus be desirable despite the information loss—political considerations are outside of the scope of our study but merit attention in future research.

References

Aaronson, D., Barrow, L. & Sander, W. (2007). Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics* 25(1): 95-135.

AERA (American Educational Research Association). (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher*, 44(8), 448-452.

ASA (American Statistical Association). (2014). ASA statement on using value-added models for educational assessment. http://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf

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.*

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.

Brown, Emma. (2016). Technical glitches plague computer-based standardized tests nationwide. *The Washington Post* (04.14.2016).

https://www.washingtonpost.com/local/education/technical-glitches-plague-computer-based-standardized-tests-nationwide/2016/04/13/21178c7e-019c-11e6-9203-7b8670959b88_story.html?utm_term=.edce982aa6a8

Brown, Emma. (2014). D.C. Public Schools Takes a Hiatus from Test-Based Teacher Evaluations as City Moves to Common Core Exams. *The Washington Post* (06.19.2014). https://www.washingtonpost.com/local/education/dc-public-schools-takes-hiatus-from-test-based-teacher-evaluations-as-city-moves-to-common-core-exams/2014/06/19/184b8b44-f7c2-11e3-8aa9-dad2ec039789_story.html

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

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, 104(9), 2633-79.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2017). Measuring the impacts of teachers: Reply. *American Economic Review* 107(6), 1685-1717.

Chin, M. (2016). A "jarring" experience? Exploring how changes to standardized tests impact teacher experience effects. *Working paper, Harvard University.*

Chute, Eleanor and Mary Niederberger. (2015). Pennsylvania Gets Waiver on Using PSSA Scores to Assess Schools, Teachers. Pittsburg Post-Gazette (09.08.2015). Retrieved on 06.21.2016 at: http://www.post-gazette.com/news/education/2015/09/08/Pennsylvania-granted-one-year-waiver-on-using-PSSA-test-scores-to-measure-school-teacher-performance/stories/201509080124

Condie, Scott, Lars Lefgren and David Sims. 2014. Teacher Heterogeneity, Value-Added and Education Policy. *Economics of Education Review* 40(1), 76-92.

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

Corcoran, S., & Goldhaber, D. (2013). Value added and its uses: Where you stand depends on where you sit. *Education*, *8*(3), 418-434.

D'Agostino, J.V., Welsh, M.E., & Corson, N.M. (2007). Instructional Sensitivity of a State's Standards-Based Assessment. *Educational Assessment,* 12(1), 1-22.

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

Dee, T. S., & Wyckoff, J. (2015). Incentives, Selection, and Teacher Performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2), 267–297.

Dimick, J.B., Staiger, D.O., Basur, O., & Birkmeyer, J.D. (2009). Composite Measures for Predicting Surgical Mortality in the Hospital. *Health Affairs*, 28(4),1189-1198.

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

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, D., & Chaplin, D. D. (2015). Assessing the "Rothstein Falsification Test": Does It Really Show Teacher Value-Added Models Are Biased? *Journal of Research on Educational Effectiveness*, 8(1), 8-34.

Goldhaber, Dan, James Cowan and Joe Walch. 2013. Is a Good Elementary Teacher Always Good? Assessing Teacher Performance Estimates Across Subjects. *Economics of Education Review* 36(1), 216-228.

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, Walch, Joe, and Gabele, Brian (2014). 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.

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*.

Kane, T.J., Taylor, E.S., Tyler, J.H., & Wooten, A.L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources* 46(3), 587-613.
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, 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*.
Lacireno-Paquet, N., Morgan, C. & Mello, D. (2014). How states use student learning objectives in teacher evaluation systems: a review of state websites. Policy Report, Institute of Education Sciences, United States Department of Education: Washington, DC.
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.
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.
McCaffrey, D.F. (2013). Will Teacher Value-Added Scores Change when Accountability Tests Change? Knowledge Brief. Stanford, CA: Carnegie Foundation for the Advancement of Teaching.
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.
NCTQ. (2016) Teacher Evaluation Policy in Tennessee. http://www.nctq.org/dmsView/Evaluation_Timeline_Brief_Tennessee
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.
Polikoff, M.S., & Porter, A.C. (2014). Instructional alignment as a measure of teacher quality. *Educational Evaluation and Policy Analysis*, 36(4), 399-416.
Rothstein, J. (2017). Measuring the impacts of teachers: Comment. *American Economic Review* 107(6), 1656-84.
Rothstein, J. (2015). Teacher quality policy when supply matters. *American Economic Review* 105(1), 100-130.
Sass, Tim R., Anastasia Semykina and Douglas N. Harris. 2014. Value-Added Models and the Measurement of Teacher Productivity. *Economics of Education Review* 38(1), 9-23.
Schmidt, W. H., & Houang, R. T. (2012). Curricular Coherence and the Common Core State Standards for Mathematics. *Educational Researcher*, 41(8), 294-308.
Schochet, P. Z., & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics*, 38(2), 142-171.
Simpson, K. and Torres, Z. (2014). Testing Portion of Colorado Teacher Evaluations Lagging Behind. *The Denver Post* (04.19.2014). Retrieved on 06.21.016 at: http://www.denverpost.com/2014/04/19/testing-portion-of-colorado-teacher-evaluations-lagging-behind/
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*.
Staiger, D.O., & Rockoff, J.E. (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives* 24(3), 97-118.
Steinberg, M., & Donaldson, M. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy*, 11(3), 340-359.
Steinberg, M., & Garrett, R. (2016). Classroom Composition and Measured Teacher Performance: What Do Teacher Observation Scores Really Measure? *Educational Evaluation and Policy Analysis*, 38(2), 293-317.
Whitehurst, G.J., Chingos, M.M., & Lindquist, K.M. (2014) Evaluating teachers with classroom observations: Lessons learned from four districts. Policy Report, Brown Center on Education Policy. Brookings Institution, Washington DC.
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*.

Table 1: *Assessment Changes During Study Period*

State	Transition	Implementation Year (Spring)	Accompanied by Standards Change?
Kentucky	1	2012	Yes
Massachusetts	1	2013	Yes
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	Year and subject	Grades	Demographic Information	Unique Teachers	Unique Students
Kentucky	2009-2014 Math and reading	4-5	Race, gender, FRPL, ELL, special education	7,577	297,347
Massachusetts	2011-2014 Math and reading	4-8	Race, gender, FRPL, ELL, special education	22,776	632,896
New York City	2000-2010 Math 2004-2010 Reading (and 2003 for grade 5)	4-8	Race, gender, FRPL, ELL, disability/special education	26,519	823,389
North Carolina	1997-2012 Math and reading	4-5	Race, gender, FRPL, disabilities	28,207	1,214,113
Washington	2006-2013 Math and reading	4-5	Race, gender, FRPL, 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 Decile Teachers in t-1 Remaining in the Same 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

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 that the value from the transition periode is significantly lower than in both surrounding stable regimes at 95% (*) and 99% (**) significance levels.

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

Kentucky

	Math			Reading		
Regime	Adv	Avg	Disadv	Adv	Avg	Disadv
Stable 1	55.8	48.9	44.1	56.8	48.9	42.8
Multitest Yr	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
Multitest Yr	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

New York City

	Elementary						Middle					
	Math			Reading			Math			Reading		
Regime	Adv	Avg	Disadv	Adv	Avg	Disadv	Adv	Avg	Disadv	Adv	Avg	Disadv
Stable 1	62.1	49.1	38.9	71.7	49.1	32.9	64.3	43.9	44.6	72.0	44.8	39.0
Multitest Yr	65.8	45.8	40.0	66.6	48.7	40.6	65.8	48.6	38.9	68.4	51.9	40.2
Stable 2	64.0	45.3	44.7	67.7	45.1	41.2	64.9	48.7	38.8	72.6	47.0	35.6

North Carolina

	Math			Reading		
Regime	Adv	Avg	Disadv	Adv	Avg	Disadv
Stable 1	55.6	50.5	48.7	59.0	49.7	45.0
Multitest Yr	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
Multitest Yr	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	Avg	Disadv	Adv	Avg	Disadv
Stable 1	61.4	48.0	44.1	60.0	47.6	46.9
Multitest Yr	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 FRPL, average ("avg.") is the middle quintile of each, and disadvantaged ("disadv.") is the bottom quintile of average prior achievement and top quartile of percent FRPL. Significance stars indicate that the value during the multitest year is significantly lower than in both surrounding stable regimes at the 95% (*) and 99% (**) significance levels.

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

	Kentucky		MA Elem		MA Middle	
	(1)	(2)	(3)	(4)	(5)	(6)
%FRPL			-0.31	-7.57**	4.31**	4.23*
			(0.92)	(2.17)	(1.16)	(1.91)
%Black	-1.90*	-3.53**	2.27*	0.96	1.09	2.20
	(0.96)	(1.13)	(1.36)	(3.06)	(1.47)	(2.57)
Premath	-0.84	-1.47	-1.20*	-4.79**	0.99	1.93
	(0.80)	(1.01)	(0.71)	(1.61)	(0.75)	(1.26)
Preread	-0.89	-1.31	1.37*	1.26	-0.04	-1.48
	(0.84)	(1.01)	(0.73)	(1.69)	(0.81)	(1.35)
Math	-2.01**	-1.60**	0.34	0.39	-0.66	-0.48
	(0.30)	(0.37)	(0.28)	(0.60)	(0.35)	(0.57)
Transition	2.50**	1.67*	1.31**	-2.34*	0.36	0.90
	(0.33)	(0.67)	(0.37)	(1.01)	(0.33)	(0.91)
Transition * Math		-1.02		-0.04		-0.26
		(0.59)		(0.65)		(0.66)
Transition * %FRPL				8.49**		0.09
				(2.31)		(2.15)
Transition * %Black		3.82*		1.28		-1.62
		(1.79)		(3.35)		(2.91)
Transition * premath		1.38		4.07*		-1.37
		(1.58)		(1.76)		(1.53)
Transition * preread		1.13		0.21		2.10
		(1.66)		(1.82)		(1.58)
Constant	25.66**	25.16**	20.27**	23.46**	17.18**	16.87**
	(0.31)	(0.44)	(0.52)	(0.93)	(0.53)	(0.79)
Observations	17482	17482	18227	18227	12994	12994
R-squared	0.007	0.010	0.003	0.005	0.006	0.007

Notes: The 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.
* Significant at 5% level ** Significant at 1% level

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

	NYC Elem		NYC Middle		North Carolina		Washington	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
%FRPL	0.15	0.395	0.07	0.20	2.70*	1.97*	0.46	1.11
	(0.38)	(0.47)	(0.54)	(0.66)	(0.42)	(0.49)	(0.65)	(0.78)
%Black	0.59*	0.409	0.69	0.26	0.81	1.18*	0.64	0.18
	(0.33)	(0.36)	(0.42)	(0.49)	(0.39)	(0.45)	(1.47)	(1.80)
Premath	0.47	0.23	0.08	-0.27	-0.93*	-1.32*	0.98	0.54
	(0.43)	(0.48)	(0.57)	(0.64)	(0.33)	(0.38)	(0.53)	(0.64)
Preread	-1.17*	-1.00	-1.94*	-1.78*	0.226	0.07	0.13	0.66
	(0.41)	(0.46)	(0.57)	(0.64)	(0.36)	(0.41)	(0.59)	(0.71)
Math	-0.16	-0.10	-4.58*	-4.86*	-4.54*	-4.09*	-2.20*	-2.21*
	(0.17)	(0.21)	(0.23)	(0.30)	(0.12)	(0.18)	(0.22)	(0.26)
Transition	0.44	-0.22	-0.66*	-2.34*	-0.55*	0.57	0.33	2.06*
	(0.21)	(0.84)	(0.24)	(1.15)	(0.13)	(0.46)	(0.25)	(0.72)
Transition * Math		-0.62		0.32		-1.55*		0.02
		(0.36)		(0.47)		(0.45)		(0.47)
Transition * %FRPL		-0.26		-0.08		-0.95		-2.06
		(0.85)		(1.17)		(0.79)		(1.29)
Transition * %Black		-0.01		1.52*		0.62		1.39
		(0.76)		(0.86)		(0.67)		(3.01)
Transition * premath		0.05		1.53		1.02		1.31
		(1.03)		(1.32)		(0.64)		(1.05)
Transition * preread		-0.09		-0.83		-0.68		-1.62
		(0.98)		(1.31)		(0.69)		(1.24)
Constant	23.15**	24.25**	24.55**	25.69**	26.27**	25.82**	24.27**	23.74**
	(0.36)	(0.57)	(0.52)	(0.78)	(0.22)	(0.33)	(0.38)	(0.44)
Observations	57481	57481	32688	32688	113213	113213	31096	31096
R-squared	0.001	0.002	0.019	0.02	0.015	0.018	0.004	0.005

Notes: The 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.
* Significant at 5% level ** Significant at 1% level

Table 6. *Out-of-Sample Forecasts of Value-added*

	Single-test-year forecasts (pooled)	Multitest-year forecasts (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)	
Massachusetts elementary						
	1.010	0.899**	0.901**		0.897**	
	(0.016)	(0.016)	(0.022)		(0.019)	
Massachusetts middle						
	1.044**	1.057**	1.100**		1.008*	
	(0.014)	(0.019)	(0.026)		(0.024)	
New York City elementary						
	1.023**	1.068**	1.089**		1.039	
	(0.008)	(0.021)	(0.025)		(0.028)	
New York City middle						
	1.042**	1.107**	1.122**		1.081*	
	(0.010)	(0.022)	(0.025)		(0.040)	
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)
Washington						
	0.973*	0.933**	0.984		0.841**	
	(0.011)	(0.022)	(0.027)		(0.028)	

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. Significance stars are for a test of whether the forecasting coefficient can be distinguished from 1.0. Standard errors clustered at the school-cohort level shown in paretheses.
* Significant at 5% level ** Significant at 1% level

Notes for Figures 3a–3g: Vertical lines denote multitest years.

Figure 3a: Adjacent-Year Correlations in Kentucky, Elementary

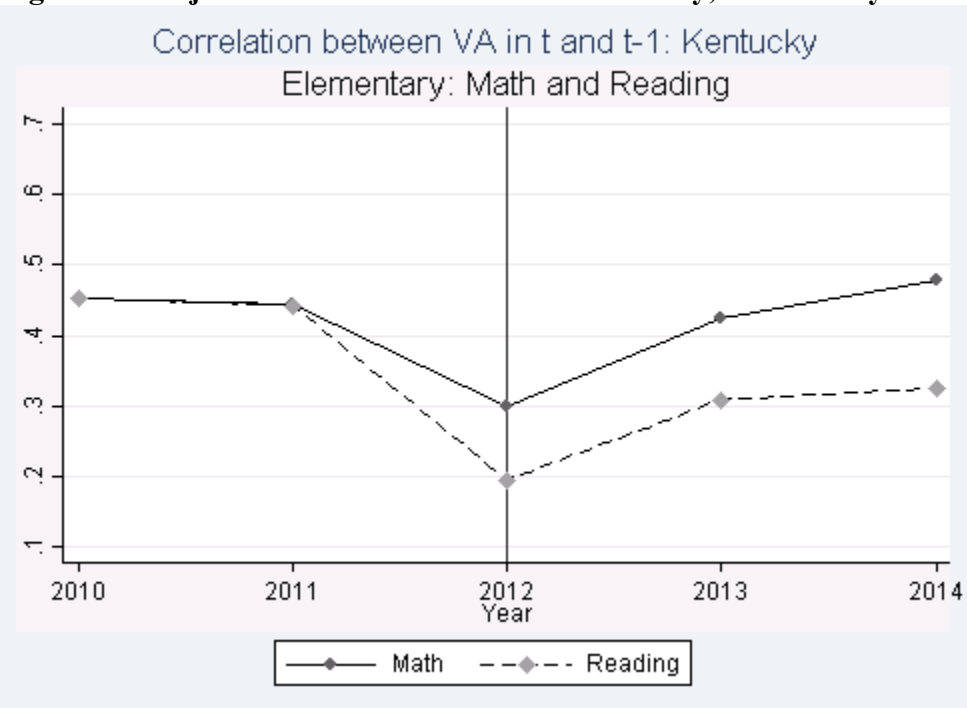


Figure 3b: Adjacent-Year Correlations in Massachusetts, Elementary

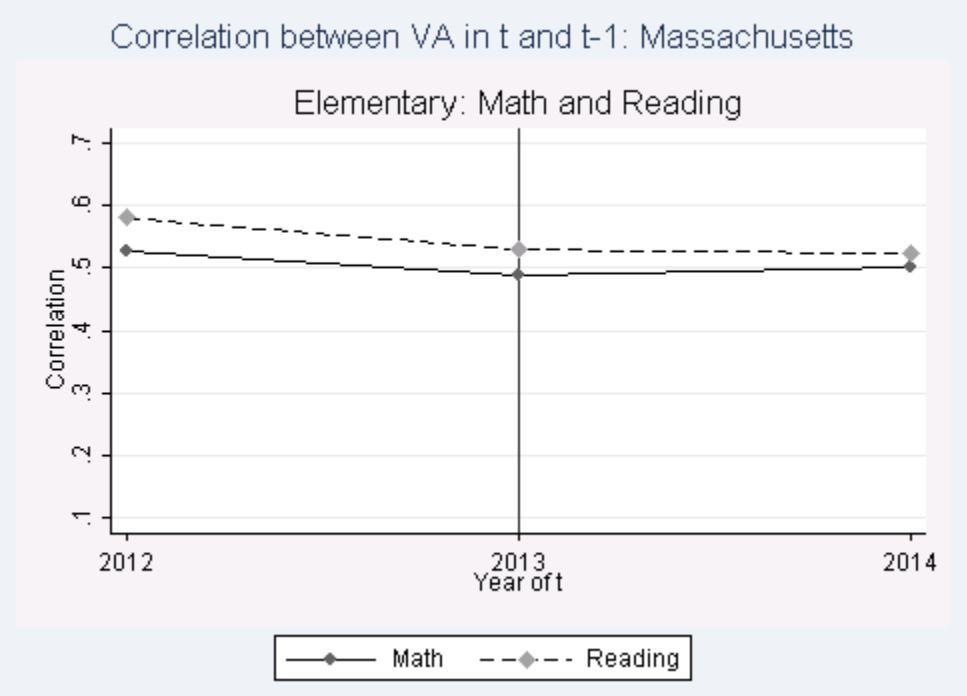


Figure 3c: Adjacent-Year Correlations in Massachusetts, Middle

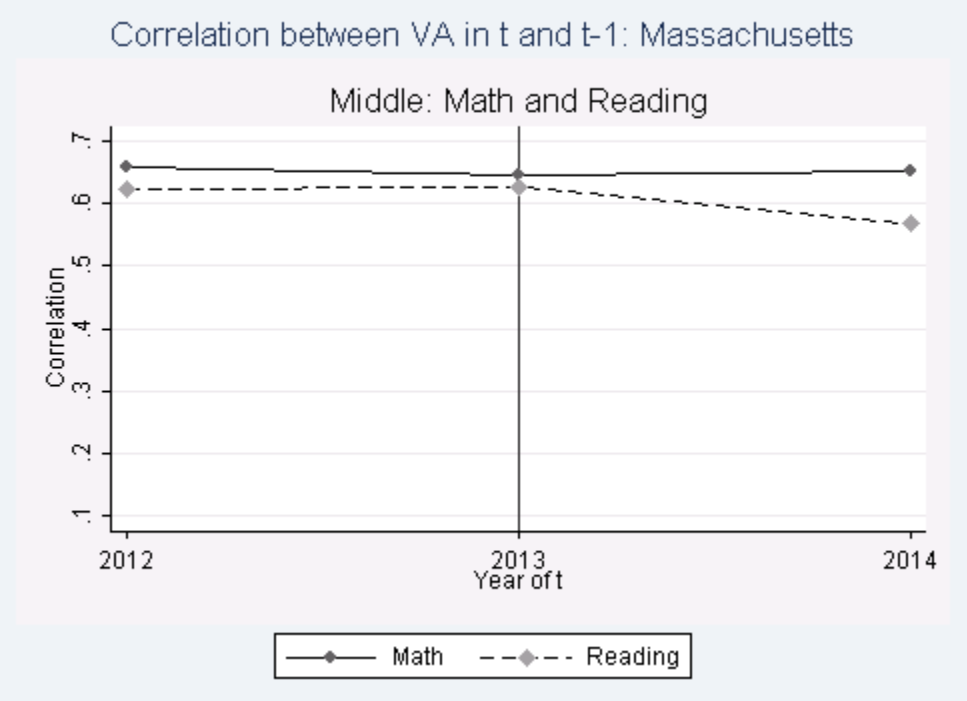


Figure 3d: Adjacent-Year Correlations in New York City, Elementary

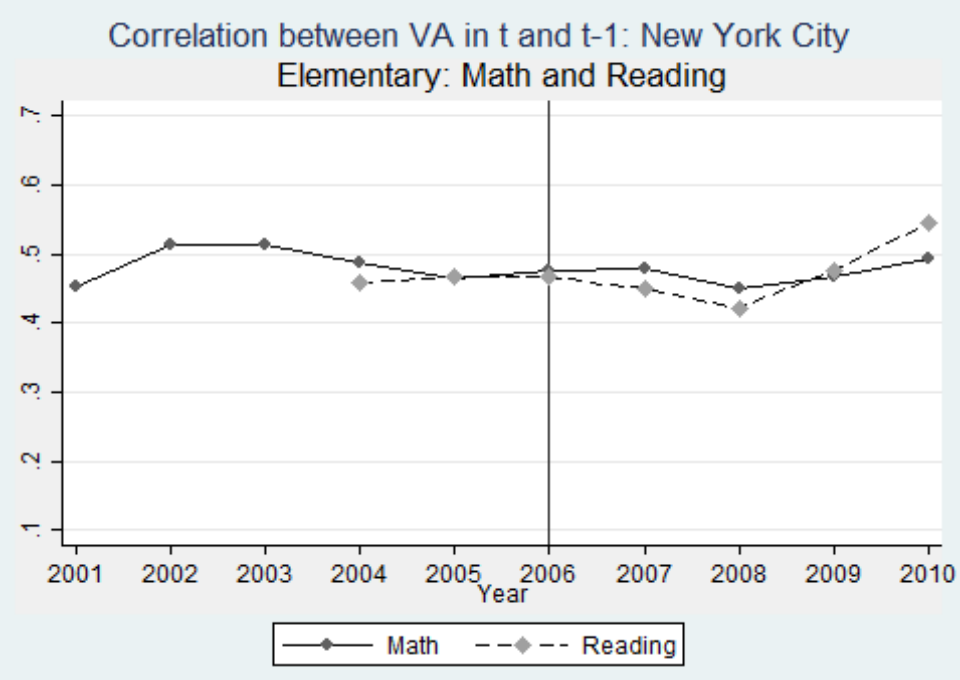


Figure 3e: Adjacent-Year Correlations in New York City, Middle

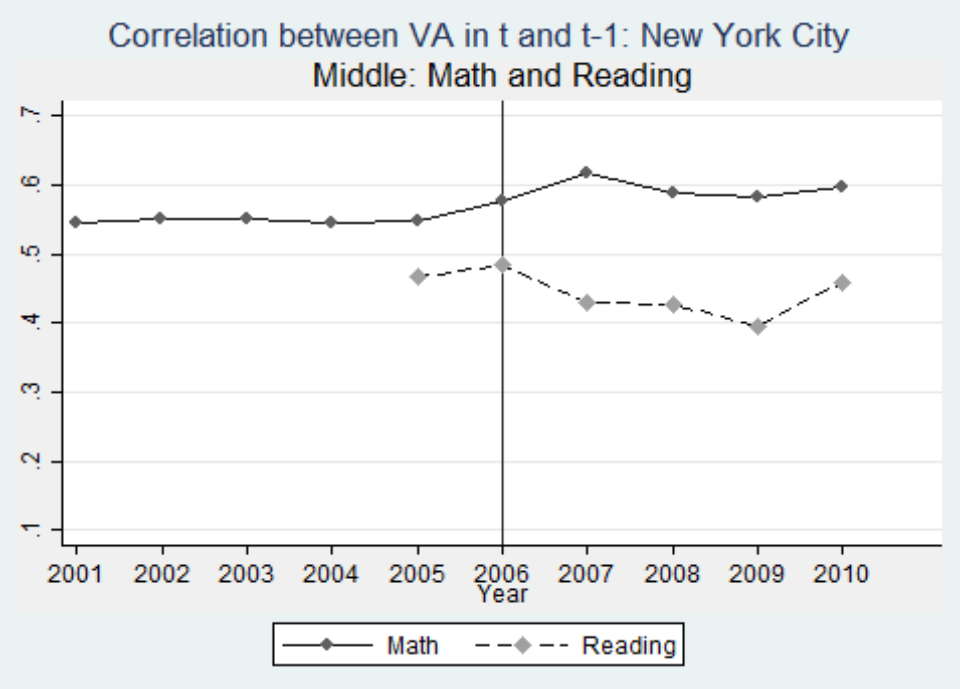
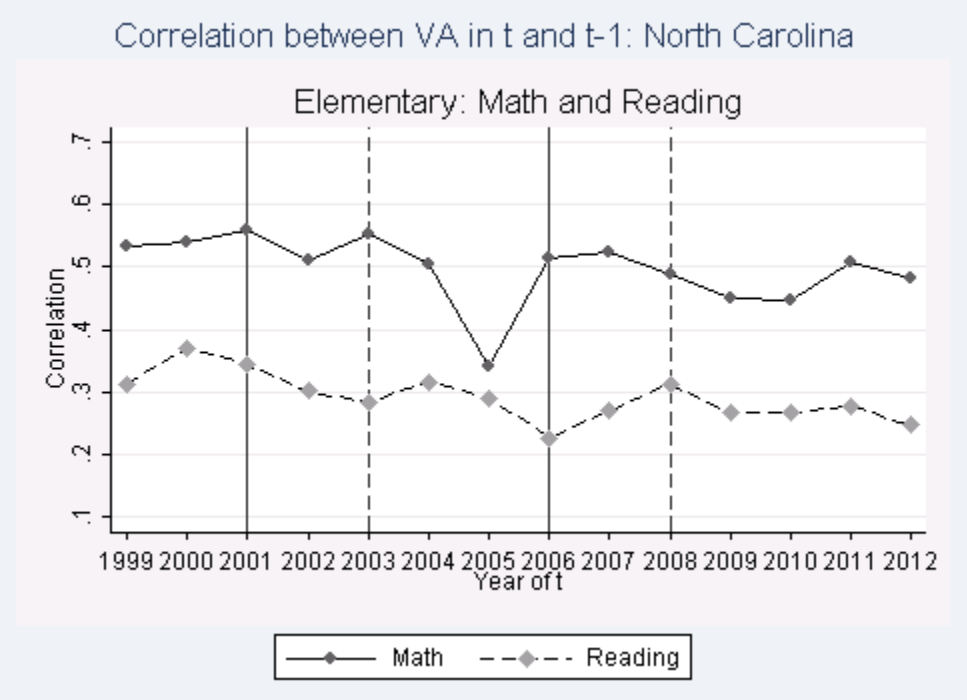


Figure 3f: Adjacent-Year Correlations in North Carolina, Elementary



Note: solid lines indicate math transitions and dashed lines indicate reading multitest years.

Figure 3g: Adjacent-Year Correlations in Washington, Elementary

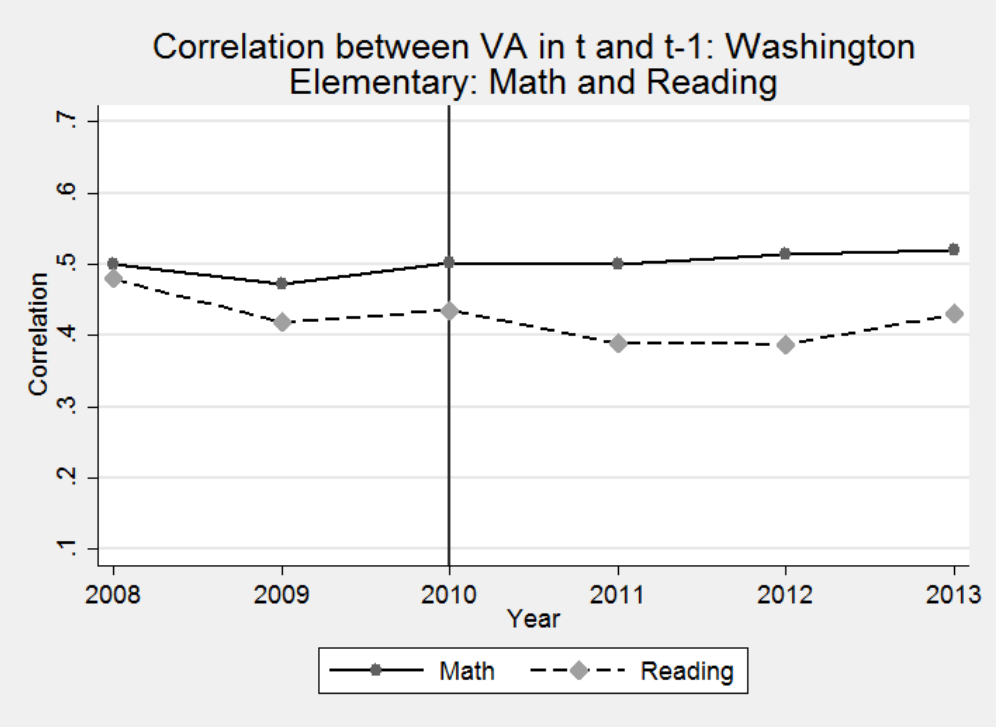
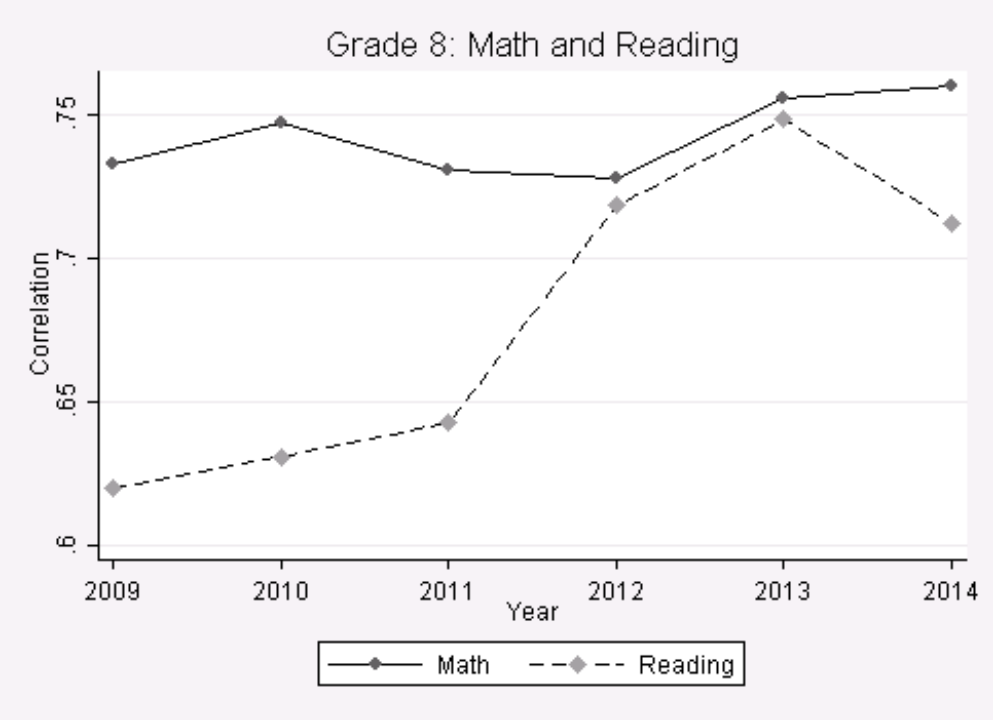


Figure 4: Student-level Correlation Between EXPLORE and EOG in Kentucky by Year



Notes: Student-level correlation between end-of-grade tests, which changed in 2012, and Explore tests (no change) in Kentucky by year.

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. As an alternate specification, we obtained results after following the two-step transformation procedure 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. We implement the Koretz et al. (2014) procedure by first dropping all 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 run the remaining test scores through the probit function and standardize the resulting scores by year, grade, and subject. Results are similar when we perform our analysis 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, subject, and school, 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.

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 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. 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.[42] Since 2010,

[42] Some of the data are linked using the statewide assessment's "teacher of record assignment" – i.e., the 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 roughly 95 percent of students in both math and reading.

Washington has directly linked students to their teacher of record. We therefore use this data to link students to their math and reading teachers in Grades 4 and 5.

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

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

To estimate correlations in true teacher effectiveness in the absence of sampling 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 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.