Economics of Education Review 62 (2018) 48–65

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

The common core conundrum: To what extent should we worry that changes to assessments will affect test-based measures of teacher performance?☆

Ben Backes[a,*], James Cowan[a], Dan Goldhaber[a], Cory Koedel[b], Luke C. Miller[c], Zeyu Xu[a]

[a] American Institutes for Research
[b] University of Missouri
[c] University of Virginia

A B S T R A C T

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.

1. 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, McCaffrey, Miller, & Staiger, 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, Taylor, Tyler, & Wooten, 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.

☆ 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.

* Corresponding author.
E-mail addresses: bbackes@air.org (B. Backes), jcowan@air.org (J. Cowan), dgoldhaber@air.org (D. Goldhaber), koedelc@missouri.edu (C. Koedel), lcm7t@virginia.edu (L.C. Miller), zxu@air.org (Z. Xu).

http://dx.doi.org/10.1016/j.econedurev.2017.10.004
Received 12 March 2017; Received in revised form 6 October 2017; Accepted 10 October 2017
Available online 24 October 2017

Teacher performance evaluations that incorporate value-added have spread rapidly in recent years.[1] 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.[2] 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, Amrein-Beardsley, Haertel, & Rothstein, 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 to the controversy.[3] 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.[4] Some policy makers and practitioners, and most prominently teachers' unions, have argued that teachers need more time to develop lessons and learn about the new tests before being evaluated.[5] A related concern is that the curricular and testing transitions did not always occur simultaneously, creating potential misalignment between the curriculum and assessment.[6] 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.[7]

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.[8] 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 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.[9] 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 explore the reliability and stability of teacher value-added during changes in assessment regimes.[10] 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.[11] 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.

We also apply the methods of Chetty et al. (2014a) and Bacher-Hicks et al. (2014) to improve our understanding of the informational content of value-added during assessment changes. Specifically, we

[1] 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).

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

[3] 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.

[4] For example, see Chang, K.. (September 3, 2013). With Common Core, fewer topics but covered more rigorously. *The New York Times*, D2.

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

[6] 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).

[7] 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

[8] 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).

[9] 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).

[10] Our analysis is along the lines of what is advocated by McCaffrey (2013).

[11] 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.